

AD-F630038

②

AD A132423

A SYSTEMATIC APPROACH
TO HUMAN FACTORS MEASUREMENT

Reproduced From
Best Available Copy

DAVID MEISTER

OCTOBER 1978

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION
UNLIMITED

20000802028

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152

DTIC
SEP 08 1983
E

DTIC FILE COPY

83 09 08 038

NOTE

This material is intended for use in connection with the tutorial session, "A System Methodology for Behavioral Research," presented at the annual meeting of the Human Factors Society, October 16-19, 1978, Detroit, Michigan. The opinions and assertions contained herein are those of the writer and are not to be construed as reflecting the views of the Navy Department or naval service.

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

DTIC
COPY
REPRODUCTION

TABLE OF CONTENTS

	PAGE
INTRODUCTION AND OVERVIEW	1
CHAPTER ONE: SYSTEM DEFINITIONS AND ASSUMPTIONS	9
CHAPTER TWO: IMPLICATIONS OF THE SYSTEM CONCEPT	21
CHAPTER THREE: CHARACTERISTICS OF PERSONNEL SUBSYSTEM MEASUREMENT	35
CHAPTER FOUR: PMS MEASUREMENT QUESTIONS	45
CHAPTER FIVE: SYSTEM THEORY	93
CHAPTER SIX: CRITERIA FOR SELECTING AND EVALUATING PSM RESEARCH	109
REFERENCES	119
APPENDIX I: CRITICISMS OF THE SYSTEM APPROACH TO MEASUREMENT	A-0
APPENDIX II: SOME REPRESENTATIVE PSM STUDIES	B-0
APPENDIX III: THE EFFECT OF GOVERNMENT ON HUMAN FACTORS RESEARCH	C-0
AND DEVELOPMENT	

INTRODUCTION AND OVERVIEW

The following points will be emphasized in this monograph. Details may be found in the chapters as noted.

Chapter One - SYSTEM DEFINITIONS AND ASSUMPTIONS

1. The further development of Human Factors requires continuing examination of its basic concepts.
2. There is a discordance between the products of Human Factors research and their effective application in system development and operation.
3. The system concept is the heart of our theoretical structure.
4. Human Factors is a system-oriented discipline because it is the study and application of factors affecting personnel performance in manned systems.
5. A system is an organization in which the individual elements work together purposefully to produce an output which the individual element can not produce by itself.
6. The following assumptions follow from the system concept:
 - a. Systems are organized hierarchically.
 - b. Systems are purposeful.
 - c. Each system element subordinates itself to the system purpose.
 - d. Each system element affects every other element.
 - e. The outputs of individual elements are transformed to produce the system output.
 - f. Measurement, evaluation and feedback are inherent in the system concept.
 - g. The system concept requires action.

Chapter Two - IMPLICATIONS OF THE SYSTEM CONCEPT

1. The operational manned system is the model for Human Factors measurement. This implies that the measurement questions we ask stem from that model. The two fundamental research questions to be answered are:

- a. What is the effect of system parameters on personnel performance?
- b. What is the effect of personnel performance on the system output?

2. The operational system is organized around the "mission scenario". Both operational and laboratory research should reproduce the essential characteristics of that scenario.

3. Research tasks must be purposeful and meaningful to subjects in terms of an actual or simulated mission goal.

4. Research studies must be validated (replicated) in or with operational systems.

5. All system-relevant factors must be included in the measurement situation. If these cannot be introduced into the laboratory, the study must be performed in the operational environment.

6. The effect on system output is the criterion of significance for personnel variables.

7. Performance must be measured on both individual and system levels.

8. The system concept emphasizes measurement in the operational environment and evaluation of performance.

9. The questions asked in laboratory research should focus on the relationship between individual and system parameters.

10. Operational equipment is not required in laboratory research as long as a mission scenario is utilized and a system output is produced by integrating/transforming outputs from several sources.

11. System simulation strongly emphasizes team situations.

12. Subjects for research studies must be highly trained for their tasks.

Chapter Three - CHARACTERISTICS OF PERSONNEL SUBSYSTEM MEASUREMENT

1. Controlled experimentation (CE) is measurement performed under highly controlled conditions found primarily in the laboratory and emphasizes manipulation of variables as part of hypothesis-testing.

2. Personnel subsystem measurement (PSM) is measurement of personnel performing the total task or job (or aspects of these) in the context of or in reference to the actual work (i.e., system) environment.

3. PSM has many more purposes than CE. PSM goals are:

- a. To determine feasibility of an approach.
- b. To select the most effective alternative.
- c. To determine capability to perform.
- d. To evaluate system and element effectiveness.
- e. To solve problems in the personnel subsystem.
- f. To perform needed research.

4. In contrast to CE, PSM in its operational measurement mode rarely manipulates variables. It emphasizes exercise of the mission scenario.

5. In contrast to CE, PSM does not need to develop hypotheses.

6. Performance criteria are inherent in operational PSM but rarely found in CE.

7. In contrast to CE, PSM takes its measurement tasks from an operational system and task fidelity is critical to it.

8. Both CE and PSM employ statistical standards of proof but PSM also requires solution of a system problem or an evaluation on which some action will be based.

9. PSM is oriented to the system, CE largely to the individual or group.
10. PSM measures at both personnel and system levels, whereas CE measures only personnel performance. PSM employs qualitative methods much more than does CE.
11. PSM employs tests and normative data gathering, the latter describing systems, differences among systems, the relation between personnel performance and system parameters, and human performance reliability.
12. PSM employs correlational analysis much more frequently than does CE, which emphasizes the testing of differences between conditions.

Chapter Four - PSM MEASUREMENT QUESTIONS

1. PSM purposes imply certain measurement questions which are organized by:
 - a. stage of system development;
 - b. training program development;
 - c. system operation;
 - d. system maintenance.
2. PSM questions related to system development include:
 - a. Do personnel possess the capability to perform certain functions to specified levels; are system design concepts feasible from a personnel performance standpoint? These questions require a type of study called an Exploratory test.
 - b. Which of two or more alternative system configurations is more effective from a personnel performance standpoint? This question demands a Resolution test.
 - c. Does the system satisfy system requirements (from a personnel standpoint)? This question requires a Verification test.

d. PSM research for this phase asks the following questions: How do system developers develop systems? How do they make use of Human Factors inputs? What is the relationship between system characteristics and operator performance?

3. Questions related to training include:

a. Has the necessary training been accomplished? Is the training adequate? (Operational study).

b. Does performance transfer from the training environment to the operational job? How does that performance transfer? (The first question requires an operational study, the second, a research study.)

c. How do the effects of training in one mode or medium compare with those of another mode or medium? (Research study).

d. How faithfully must the training environment reproduce the operational one? (Research study).

4. Questions related to system operations include:

a. How well do system personnel perform relative to requirements? (Verification test, accomplished by continuous, periodic or special system evaluation.)

b. Is the system ready to perform as required? (Operational Readiness test).

c. How can a problem arising from system verification be solved? (Investigative measurement).

d. How does a new system configuration compare with the old? (Resolution/Verification test).

5. Questions relative to maintenance include:

a. How do technicians perform diagnostic maintenance? (Research study).

b. How efficient is diagnostic maintenance? (Verification test).

Chapter Five - SYSTEM THEORY

1. System types divide themselves into military, commercial/industrial and social-benefit categories.
2. The identical elements among the system types are striking. Each contains many identical subsystems. Differences within a system type are greater than those between system types.
3. Because of this, PSM principles derived from military testing situations can be applied to the other types of systems.
4. A significant difference between military and non-military systems is that the latter provide a benefit to clients whereas military systems do not.
5. Because non-military systems involve clients, one must include as factors to be measured in those systems:
 - a. the way in which clients interact with systems;
 - b. the desires, needs and performances of clients as constraints on system performance.
6. Systems may also be described in terms of characteristics that cut across system types. These include:
 - a. types/number of functions performed;
 - b. number of operational modes;
 - c. number of subsystems;
 - d. system organization;
 - e. number/organization of operator positions;
 - f. number/type/locus of transforms;
 - g. number/organization of communications channels;
 - h. output requirements;
 - i. characteristic inputs;

- j. system reactivity;
- k. degree of mechanization;
- l. system feedback;
- m. system indeterminacy.

7. Indeterminacy is composed of three variables:

- a. the nature of stimulus inputs;
- b. the amount of flexibility permitted by procedures;
- c. degree of personnel response programming.

8. Indeterminacy has significant impact on the measurement strategy adopted.

Chapter Six - CRITERIA FOR SELECTING AND EVALUATING PSM RESEARCH

1. The criteria ordinarily applied to scientific research--validity and reliability--are not sufficient for PSM research.

2. Research criteria should be applied at two stages:

- a. Before a study is initiated, to decide whether or not to proceed with that study;
- b. After a study is completed, to evaluate the worth of the study and its results.

3. In addition to validity and reliability, PSM makes use of the following criteria: relevance; applicability; generalizability; and utility.

4. Validity in an absolute sense can never be established because it presumes a standard of comparison independent of measurement operations. It can be used only post-facto.

5. Reliability is almost never used as the basis for selecting one measurement situation over another, but it is used for evaluation. PSM has some difficulty with reliability because of the reduced control under which some PSM measurements are made.

6. Relevance indicates whether measurement results relate to the questions/ purposes for which a study was initiated. There is a hierarchy of such questions/ purposes. Relevance relates to the closeness between the specific study purpose and its higher order goal. Relevancy can be used for deciding whether to perform a study, as well as for evaluation.

7. Applicability indicates the degree to which study results can be transformed into action consequences. PSM research is usually more applicable than traditional CE.

8. Generalizability indicates the degree to which study results can describe objects or phenomena similar to but not identical to those on which measurements were made. This criterion is weaker than relevance and applicability.

9. Utility is defined in terms of three dimensions:

- a. problem criticality;
- b. whether the problem/question can be measured;
- c. whether study results can be applied in the real world.

10. Validity and reliability are quantitative (coefficients of correlation); the other criteria are purely qualitative.

Appendix I - Criticisms of the System Approach to Measurement

Appendix II - Some Representative PSM Studies

CHAPTER ONE

SYSTEM DEFINITIONS AND ASSUMPTIONS

Introduction and Purpose

This paper has been written to accompany the tutorial session on "A Systematic Approach to Behavioral Research."¹ Hopefully it will help the participant follow the organization and logic behind the ideas expressed in that session.

The key word in the preceding paragraph is "ideas." The session presents concepts, some of which are based on logic and common sense, while others are intuitive and speculative. The reader may or may not agree with these concepts; he may not even agree that they are concepts. The session is not tutorial in the sense of providing the participant with established and irrefutable facts.² Its purpose is to present these concepts, explore their implications, and see where they lead us. The intent is that participants will be stimulated into thinking about these ideas on their own and will then expand upon them.

Why were this session and this paper developed? It appears to the author that the further development of any discipline requires continuing examination of its basic concepts. Moreover, all disciplines live in two worlds--one of academically oriented research and the other of application--and this leads to some discordance. In the case of Human Factors, moreover, the discordance is

¹Presented at the Annual Meeting of the Human Factors Society, October 1978, Detroit, Michigan. The name of the paper does not correspond to the name of the tutorial session because upon reflection it seems more modest to restrict the topic to Human Factors alone. Moreover, I refer in the title to measurement rather than to research because it is necessary to distinguish between the two: Research is only a subset of measurement and what we discuss encompasses more than research.

²Although it is self-evident that Human Factors--like any other discipline--needs data, it also needs new ideas--or at least the critical re-examination of old ones--just as much. Indeed, the proposition can be advanced that without prior concepts the accumulation of data--particularly their meaningful interpretation--is impossible.

particularly marked because the products of Human Factors research³ have not been particularly helpful in advancing the state of its application. The question is why; and what can be done about it?

The System Concept

We start with the system because it is the theoretical underpinning of the structure we are attempting to erect (see Figure 1, to which we shall refer from time to time). This is because we define Human Factors as the study and application of the factors affecting personnel performance in manned systems.⁴ If that definition is correct, Human Factors is a system-oriented discipline; the behavior it deals with occurs in a system environment.

What is a system? In its most general sense a system is an organization--an arrangement--of elements in which the individual elements work together purposefully to produce an output--an effect, a product, a resultant--which the individual element could not produce by itself.

Miller (1978) defines a system as "a set of interacting units with relationships among them. The word 'set' implies that the units have some common properties. These common properties are essential if the units are to interact or have relationships. The state of each unit is constrained by, conditioned by, or dependent on the state of other units. The units are coupled. Moreover, there is at least one measure of the sum of its units which is larger than the sum of that measure of its units" (p. 16).

³A distinction should be made among the following terms: behavioral research which, encompassing all studies involving human operations, hierarchically subsumes the two following: psychological research which studies the individual and the group; and Human Factors research which deals with personnel performance in the context of manned systems. The three terms are often used interchangeably, which is unfortunate because this usage obscures important distinctions among them. The author contends that Human Factors measurement has a conceptual structure which is markedly different from that of Psychology and hence that Human Factors research which attempts to follow the psychological model is unlikely to satisfy the data needs of our discipline.

⁴Human Factors has two primary goals: (1) to determine how personnel function in manned systems; (2) to assist in the development and optimization of manned systems. The first goal describes its research function, the second its application function. However, the first has no significance without the second, nor should it be considered more important than the second.

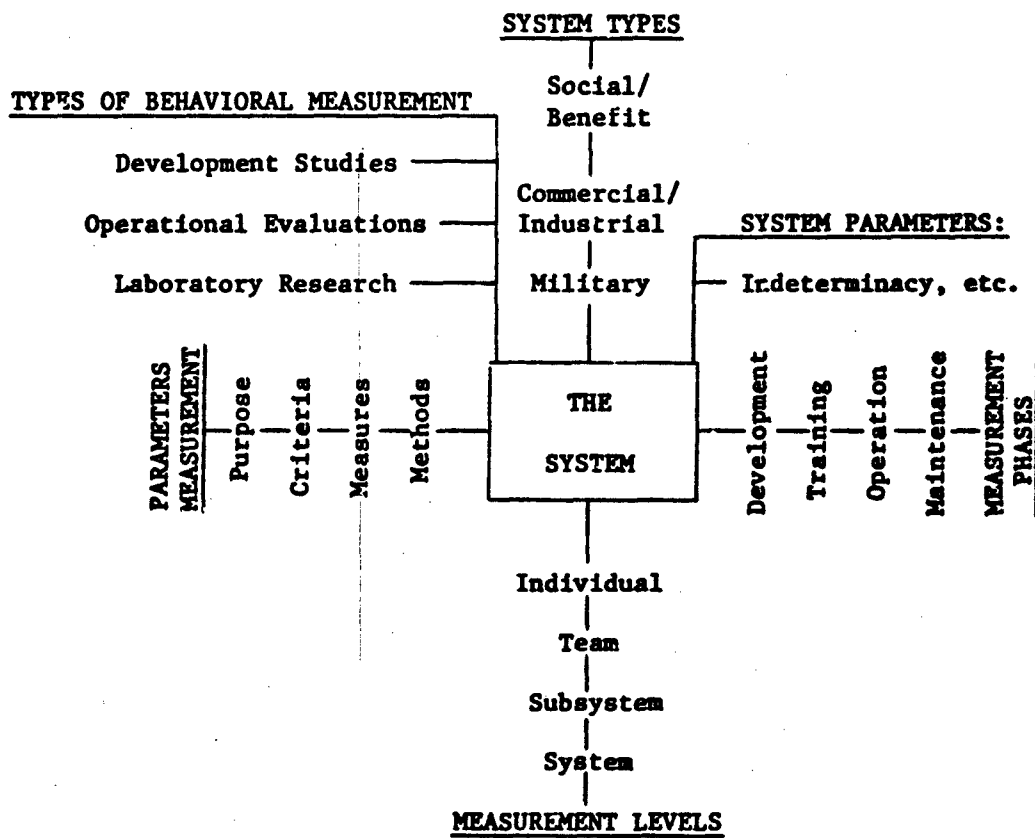


Figure 1. Dimensions of behavioral measurement.

This sounds like a restatement of the Gestalt principle, "The whole is greater than the sum of its parts," and indeed the whole--represented in our terms by the output--is in fact more than any single part. Miller points out (Miller, 1978, Note 44, p. 44) that, "Gestalt theory has had an important influence on current system theory."

There are many systems: conceptual, mathematical, social, biological, chemical physical, etc. Our concern is with manned systems, a term the author prefers to the more traditional term, man-machine systems, because the latter suggests an exclusive emphasis on mechanization. Manned systems may vary in terms of the number and sophistication of the machines they employ, some systems using only a few simple machines. Despite mechanization--or lack of it--they are all manned systems and the system definition and assumptions apply to all of them.⁵

We hope to demonstrate that the system concept has significant implications for Human Factors and its measurement processes. It would be only fair, however, to point out a number of criticisms of the system concept.

1. The definition of system is so general that everything is a system. Indeed, Miller (1978) has produced a chef d'oeuvre in which he attempts to demonstrate that the same system processes are to be found in seven hierarchical levels of living systems, starting with the cell and ending with the supranational system. An all-inclusive definition has difficulty differentiating and therefore is unlikely to be productive.

⁵What is the difference between an unmechanized manned system and the kind of ad hoc groups psychologists usually study? When:

- (1) the group has a purpose or goals for which it comes together;
 - (2) the characteristics of the group have been specified by others than the group participants;
 - (3) the goals and purposes of individuals within the group are subordinated to the overall group goal;
 - (4) procedures for implementing the overall goal are specified, from which deviations are not acceptable;
 - (5) a standard of individual and group performance exists (overtly or covertly expressed) which determines group efficiency--
- the group is a manned system, even though it may use few or no machines at all. For example, the U. S. Congress is a manned system.

2. It is difficult to know exactly what this system is that we are talking about because everything is continuous: What one considers a system at one (lower) level is also a subsystem in a higher order system. In biological terms, for example, the autonomic nervous system is only one subsystem of the human body.

These are reasonable objections but they need not concern us unduly. Although at a molar level everything is a system, at a somewhat more molecular level it is possible to make meaningful differentiations. After all, both men and women are human, but one can still tell the difference between them. It is true that a system at one level is also (and at the same time) a subsystem at another level. If, however, there are rules for specifying the boundaries of whatever one wishes to consider a system, then the objection is immaterial. Indeed, the fact that one can view the same set of objects or phenomena in different ways becomes a great advantage in viewing system interrelationships.

A number of assumptions are associated with the system concept. These are:

1. Systems are organized hierarchically, which means that one system is "nested" in another.
2. Systems are purposeful. Manned systems are purposeful because they are constructions.
3. Each system element subordinates itself to the system purpose.
4. Each system element affects every other element.
5. The outputs of individual elements are transformed to produce the system output.
6. Measurement, evaluation, and feedback are inherent in the system concept.
7. The system concept demands action to modify the system when necessary.

We discuss each of these assumptions in more detail.

Systems are organized hierarchically. It is probably a cliché to say that there are no discontinuities in life; that is why, for example, we can hypothesize a relationship between the atoms in protoplasm and a "higher order" phenomenon such as "cognition." If we think of protoplasmic atomic structure as being a system and cognition as also being a system (although obviously not on the same level), it is obvious that one system "nests" into another, and the snugness of this nesting often makes it difficult to pinpoint where one system leaves off and another begins. Probably there are intermediate systems between protoplasm and cognition, but this does not destroy the assumption; it merely extends the hierarchy.

The importance of the hierarchical assumption and the concept of "nesting" is that it enables one to note relationships between apparently disparate objects and events. Without such an assumption it would be impossible to specify the impact of an individual human response on the output of a much more complex system.

How does one define where one system leaves off and another begins? One could say that this is merely a matter of the observer's choice, but such an answer is not very satisfactory. When we say that a lower order system is nested in another higher order system, we mean that the outputs of the lower order system are received by the higher order one and are transformed in the process. (We shall discuss transforms later.) To determine the boundaries of any system it is necessary to look at the outputs of that system as they are utilized by another system.

The following example illustrates the principles described. A sonar system receives electronic returns (outputs) which are initially interpreted by the sonarman (first subsystem) as pips on a CRT, then transformed by him into a classification of submarine with a certain bearing, range, and depth. This information is passed on to the Combat Information Center (CIC) where the plotter (second subsystem) plots successive bearings, ranges, and depths onto a track. The CIC Officer (third subsystem) analyzes these data (together with voice reports from the sonarman) and decides on an attack strategy which is communicated to the Weapons Officer (fourth subsystem?) in the form of commands. Note that each subsystem receives outputs from another subsystem which may be lower order (as the sonarman is to the CIC Officer) or parallel (at the same level). In each case these outputs are received by the subsystem and transformed into something

else (CRT pipe into a submarine classification; bearings, ranges, and depths into a target track; target track into an attack strategy; the order to fire weapons into weapons adjustments).

Systems are purposeful. An academic Psychology influenced by Behaviorism is likely to "pooh-pooh" the notion of purpose (at least it did when the author went to school) because it is difficult to see purpose in individual molecular operations. However, the notion of purpose is critical to the manned system, because the manned system is a construction, i.e., it is artificial and was created according to the will of its developers. It may be difficult to discern purpose in what an individual does (although he often says he does so and so for such and such reasons); but it is obvious that a system which does not normally exist in nature was developed to serve some purpose, however obscure. Where purposes in manned systems are obscure, it is because system developers often do not think logically.

System purpose or goal is critically important for several reasons.

1. It is the starting point for the development and analysis of the system (i.e., function allocation).
2. It directs the performance of system personnel.
3. It permits the system manager to determine whether the system performs correctly.

It is reason 3 that is most important to the measurement specialist. The purpose when specified in quantitative terms becomes the standard against which performance can be judged. Unlike much individual behavior, which lacks a standard of correct performance, the system when properly designed has a standard built into it. This becomes a sort of warning indicator for major deviations from requirements.⁶

⁶ Is it possible to categorize systems in terms of their performance standards? For example, standards may vary in terms of the following: the range of what is acceptable performance; whether the standard includes qualitative (e.g., "smoothness") as well as quantitative factors (e.g., speed of response); whether the standard requires the system to maximize its performance (to output as much as possible) or to optimize it (output some level less than maximal). Each of these aspects influences the nature of the measurement strategy pursued.

The purpose also provides the action-orientation inherent in the system concept. Since the system was developed to perform at a specified achievement level, the accomplishment of the system goal is all-important to the developer and manager. Failure of the system to achieve its goal(s) means first that something is wrong and secondly that something must be done to correct the deficiency to bring the system up to standard.

Each system element subordinates itself to the system purpose. Obviously the system cannot channel all its element outputs to the common goal unless these elements subordinate themselves to an overall requirement. What this means for the human is that he follows procedures set up to implement the system goal; that he works as hard and as fast as those procedures demand, and if he fails at his job he is reprimanded, retrained, or fired.

The reader may have a mental picture that we view system personnel as if they had no will, goals, or desires of their own; that the picture we are presenting is a totalitarian one. It is true in the system context that all these individual desires and idiosyncratic behaviors are so many irrelevancies to the system goal and potentially harmful to system output--if they are not carefully controlled. The greatest disruption occurs when system personnel refuse to behave as system personnel, i.e., they strike or quit. However, the picture is not so one-sided, because if the system developer requires more of his personnel than they can produce, the system will fail (a partial justification of Human Factors is to avoid and prevent such situations). And of course the individual worker does have a certain degree of freedom; if he is working on a non-military system he can in the last resort strike or quit, daydream inefficiently or do something else undesirable (from a system standpoint). So there is play between the system and the individual, but only at the outer limits of what one can ask of the individual. One cannot require him to work under conditions of unacceptable hazard or discomfort, for example; but within those extreme limits the individual operator must conform to the system. The system operator functions at one level as an individual, on another as a system element. These two functions are exercised by the individual concurrently. How much he functions as an individual, how much as a system element, depends on how the system is programmed. The more rigidly system procedures are established, the more the operator is programmed to function as a system element; if system procedures are contingent and responses are controlled at the operator's volition, the more freedom he has. (See also the later discussion on system indeterminacy.)

Each system element affects every other element. Because all system elements interact, they exercise an effect on each other. Of course, the amount of that effect varies. For example, the internal componentry of an equipment does not affect the equipment operator's performance directly unless the equipment malfunctions. However, the way in which the circuits are designed may increase equipment capability (i.e., detection range) which does affect the operator's performance. The effect may be indirect and subtle but significant nonetheless.

The researcher has the responsibility to ascertain the amount of interaction and its resultant effect, which obviously varies with various conditions. The way in which that responsibility is usually carried out is by means of hypothesis-testing, the purpose of which is to determine whether two or more variables are significantly related. The experimental method is adequate to this task provided that all the variables found in the operational situation are permitted to exercise their effect, however small. The problem lies in the control which is considered necessary for hypothesis-testing; by eliminating or controlling all variables other than those considered by the experimenter to be immediately relevant, the experimenter "purifies" his design but also renders it artificial, since his measurement situation violates the principle with which this subsection began.

If two variables have been found by experimental test to be significantly related to each other, does this tell us any more than what we assume as a consequence of the system concept? A statistically significant "t" or "F" merely tells us that we can assume with greater or less confidence that the two variables are related in the real world. It does not tell us the amount of relationship, although regression analysis is helpful in that respect. Even a regression analysis does not help very much, however, because the experimental design has created a "purified" measurement situation, so that impact is either grossly exaggerated or minimized.

Individual system elements are transformed to produce the system output. The essence of the system is that the output of an individual element is transformed (along with the outputs of other system elements) into the system output. The logic of the system concept requires that the system output differ from any element output, because if the two were identical, the latter would in fact be the system output and thus the element alone would be the system—which would be a contradiction in terms.

Transformations within the system are therefore necessary. We are familiar with such transformations in the physical sphere, e.g., matter into energy, and they occur also in an individual behavioral context. An individual behavioral transform is the modification of an input or output through behavioral mechanisms (which may however be machine or computer-aided) into some form other than that in which it is received or output by the operator. On a physiological level, for example, a transform occurs when light stimuli are acted upon by the visual cortex to become the perception of physical shapes.

System behavioral transforms are, however, at a different level from individual behavioral ones; the former are overt and much more molar; they can be recognized without the aid of sophisticated instrumentation. Individual behavioral transforms often serve to implement system behavioral ones; for example, recognition of a visual shadow on a sonar CRT as having a particular shape leads to the judgment "this is a mine."

System behavioral transforms are more often perceptual and cognitive rather than psychomotor or motor, although obviously molecular physical processes are changing their form, as, for example, when a pilot transforms neural energy into kinetic by increasing or decreasing throttle. At least two forms of system behavioral transforms can be identified:

1. Information is coded or recoded. This information is overtly recognizable as information; it is overtly and deliberately changed by applying cognitive processes, e.g., analysis. For example, the decoding of a cyphered message.

2. A decision is made on the basis of one or more inputs. The decision is recognizable as a deliberate choice among alternatives; where the choice is lacking, as in very stable habits, there is no decision. The inputs have been transformed as a result of applying certain decision criteria or rules to those inputs.

Transforms are necessary, but why are they interesting to the system analyst?

1. By their nature transforms are critical to system performance. Consequently errors or inadequacies occurring in these transforms have special significance because they represent what one might call "fracture points," points of weakness, for the system structure.

2. Transforms can occur in several ways:

- a. Within the individual operator, when he operates on his own stimuli as in coding or decoding messages.
- b. Between team members, as when they communicate inputs or information to each other.
- c. Between subsystems, in communicating information or other outputs, to be processed or transformed by the receiving subsystems.

From the above one can see why an examination of communications channels is so important in system analysis. What is transformed behaviorally is in most cases information passed along those channels, although obviously in production systems, hospitals, etc., behavioral system transforms are paralleled by physical ones.

Depending on the complexity of the system organization, behavioral transforms may occur in series or in parallel before the final transform, the system output, is accomplished. The complexity of a system organization is represented not only by the arrangement of its physical elements and its communications channels, but also by the number of its transforms and the way in which these are accomplished.

It must be recognized that the concept of system behavioral transforms is still rather tentative at the moment and requires much more thinking.

Measurement, evaluation, and feedback are inherent in the system concept.

For the validity of this statement we must refer to a previous assumption: that systems are purposeful. If the purpose has been specified--and particularly if it has been specified quantitatively--that purpose establishes a standard of performance which must be accomplished if the system is to be said to be performing effectively. (Later we shall say more about the performance standard, its relationship to the criterion, and the differences between system-oriented and laboratory measurement as these differences relate to the criterion/performance standard difference.)

If the system has, so to speak, a "job to do," then it is logical to ask whether it is performing that job. Hence measurement and even more, evaluation,

are logical, inherent in the system concept. Measurement is necessary to determine if the performance standard is being met by the system. If we are dealing with a system which is not yet fully operational, or which only intermittently performs its functions, then measurement is necessary to determine if the performance standard can be met.

If the system is not performing adequately (i.e., the performance standard is not satisfied), then that information must be fed back to those who manage the system so that action can be taken to modify system operations in the desired direction. Hence feedback mechanisms are also essential in the development and operation of the system.

Measurement, evaluation, and feedback are essential to any properly designed system; if mechanisms are not provided to implement these functions, the system has been poorly designed. There is a parallel here between the manned system and the human system, since homeostatic sensing mechanisms are built into the human body to enable it to survive. In the manned system, however, they must be consciously designed into the system structure and measurement must be deliberately performed. In many operational systems, there is less measurement than is needed because system developers and managers have an inadequate understanding of the importance of measurement to proper system functioning. Lacking measurement processes, the system may "drift off its standard," thereby reducing the effectiveness of the mission.

The system concept demands action to modify the system when necessary to achieve the standard. If measurement indicates that the system is not performing effectively, some action must be taken to modify its design or operations. In human terms, the system is "ill" and must be restored to health. The problem must be investigated and therapeutic measures taken. If the performance standard is meaningful, the system must be brought back into accordance with that standard.

Because the system is an artificial construction, because the subject matter of Human Factors is that system, Human Factors has a responsibility to assist in the development of the system and later in the optimization of its operations. Therefore, its measurement functions, which encompass much more than research, begin in the early stages of system development and extend throughout the system life cycle. Because of its action-orientation Human Factors measurement, unlike more closely circumscribed research, is a prelude to some action; and that anticipated action in fact suggests the questions which the Human Factors measurement deals with.

CHAPTER TWO

IMPLICATIONS OF THE SYSTEM CONCEPT

Some of the implications of the system concept for Human Factors measurement were anticipated previously, but here we consider them in greater detail.

Implications for Measurement in General

The operational manned system is the model for the measurement situation. The operational manned system performing its operational (i.e., real world) functions must be the model for Human Factors measurement because Human Factors is oriented around that system.

What does it mean to say that the system is the model for our measurement situation? There are three ways in which that model affects measurement:

1. To begin at the very beginning of the measurement process, the questions we ask stem from that model. The two most fundamental research questions to be answered are:

- a. What is the effect of system parameters on personnel performance?
- b. What is the effect of personnel performance on system output?

These and more specific measurement questions which derive from them will be considered in greater detail later. Since the manned system has parameters peculiar to itself (i.e., not found in individuals, although obviously there is some overlap), the system model described in Chapter One requires the investigator to ask certain questions which he would not ask if that model did not exist. His research, therefore, is (or at least should be) profoundly influenced by system definitions and assumptions.

2. The operational system is organized around what can be termed the "mission scenario," i.e., a purposeful sequence of tasks starting with a well defined initial stimulus and progressing to an end-point defined in terms of an overall system goal. Individual tasks are integrated into the total sequence

and have little meaning apart from the overall goal. Every manned system (civilian and military) has a mission scenario. As an example of a mission scenario, the U. S. Postal Service receives letters, processes and sorts them both manually and with equipment aid, ships them, delivers them--each function and task being performed according to carefully specified procedures.

The implication of point 2. for research in both the operational and laboratory environment is that the measurement situation should endeavor to reproduce the essential characteristics of the mission scenario as the directing force for subject activity. (Obviously this is much easier in the operational environment.) The major requirements are that:

1. Tasks must be purposeful in terms of a larger goal, i.e., they must have meaning not only in and of themselves, but also in terms of implementing the specified system goal. Part-tasks or subfunctions like simple reaction time are not admissible in system-oriented research unless one is studying task mechanisms.

2. The significance of these tasks must be meaningful to the subject, not necessarily in terms of his own personal interests, but in terms of what he understands the system goal to be.

For example, if the investigator asks the subject to count all the patterns he finds in a dot mosaic, the latter might be told that (a) the mosaic represents a new type of detection display; (b) that the ability of humans to use such displays for detection of underwater targets is presently very questionable; (c) that each pattern found can be classified in terms of distinctive types of enemy mines; (d) that he should work as fast and accurately as possible because his information must be passed to another member of the team who will make a tactical judgment based on his information.

Although none of this may be strictly the truth, it does help to make the task more meaningful to the subject.

To introduce this "realism" into the experimental situation will undoubtedly create some difficulties for the investigator seeking to conduct laboratory research because it complicates that laboratory situation. Nonetheless, despite

the added effort, it is possible to incorporate mission characteristics into laboratory research.

3. To skip to the other end of the measurement process, in whatever environment we conduct our research, studies must be replicated (validated) by measuring the same phenomena in or with operational systems in the operational environment. (Or at the very least by some approximation of the operational system; there will, of course, be times when it is impossible to use the operational system for validation purposes.) If validation is defined as testing measurement conclusions reached in the laboratory against reality, that reality for the Human Factors researcher must be the operational system. Validation in the operational environment presents certain difficulties that must be recognized, not the least of which is exposing one's conclusions to risk in an environment much more complex and less well controlled than the one from which these conclusions came originally; but these difficulties can be overcome.

This form of validation is very unlikely to be implemented by many researchers, particularly academicians who have little acquaintance with, love of and opportunity to secure operational systems as testbeds. Perhaps this burden should be assigned to researchers in government laboratories. Nevertheless, the continuing failure to validate behavioral conclusions in the real world (the operational environment) means that most of our data and conclusions are suspect; they may not be invalid, but (just as bad) they may be irrelevant to reality. All system-relevant factors must be included in the measurement situation. The laboratory researcher endeavors to include--or at least control--all the factors he feels will influence his subject's performance. However, because he deals only with individual performance, he includes only those factors relevant to that performance; obviously, if what one is attempting to measure is very molecular, system factors will be largely irrelevant. Psychophysical studies, e.g., determination of minimal visual angles, do not require a system context (although they should be validated operationally).

The system-oriented investigator endeavors to include not only individual factors but those affecting the output of the actual or simulated system. Consequently the number of variables to be included in the system-oriented measurement situation increases--which does not make it any easier for the investigator.

What are these system-relevant factors? They involve two aspects:

1. All those variables that would in the operational environment be expected to affect system output. The specific nature of these variables will depend on the type of system being measured or (in a laboratory situation) simulated. For example, if the researcher develops a measurement situation based on an information-processing system model, he might include: number and type of information channels; types of messages; message frequency and familiarity, etc. If he bases his measurement on a visual surveillance system model he might include: types of stimuli; their intensity, complexity, classification rules, etc.

2. The second aspect of system-related factors is the amount of interaction the researcher includes in his system representation.

Obviously in the operational environment he does not have to pick and choose among variables to be included in the measurement; they already exist in that situation; and all he has to do is to ensure that those variables truly represent the system as an operational entity: that the system is performing under normal operational conditions using only operational procedures, etc. If the system being studied is exercised under non-normal conditions, the variables present in the measurement situation are non-representative and will produce invalid data.

For the researcher working in a more controlled environment (e.g., the laboratory) if the tasks presented to subjects are modelled directly on those of an operational system, then the major characteristics of that system must be deliberately included in the measurement. If, as is more common, synthetic tasks are developed to create what might be called a "system analogue," it is necessary for the researcher to include sufficient system variables to represent the system. (Whatever the synthetic task situation developed by the researcher, implicit in what he creates as a task or tasks is an operational system model of some sort, conceptual rather than concrete perhaps, but a model nonetheless. To develop the most effective set of experimental tasks, he must therefore examine that implicit model.) This is because a system is defined in part by the interactions among its elements. If there are too few such interactions, the task representation will lose its system character. The precise amount of

interaction needed for system representation cannot be specified, but the researcher must avoid the situation in which only one variable is abstracted from the operational system and used as a model for the development of task materials. This was, for example, the situation in which Bavelas (1950) in his pioneering study of communication processes abstracted only the patterning of information channels and created a methodology based solely on the arrangements of networks, e.g., Star, Circle. This representation was so limited that the results of the many network studies he initiated have been largely sterile.

The rationale for including system-relevant factors is the assumption that all system elements affect each other and that consequently exclusion of some from the measurement would produce an aberrant set of results. From the traditional experimental design standpoint, if the investigator excludes undesirable (e.g., potentially confounding) variables from all the groups being contrasted, this cancels out the effect of these variables on his measurement situation. But if one assumes that system elements affect one another, the exclusion of the inconvenient variables changes the way in which the remaining variables behave. Exclusion creates a non-operational (and hence invalid) measurement situation. It is possible that the well known inability of much laboratory research to predict or explain operational performance (Chapanis, 1967) results from the non-operational character of that research.

To manipulate all these variables and their values (particularly in an orthogonal manner) would require a measurement situation of tremendous complexity.

A partial solution is to create a situation in which system variables are allowed to influence performance as they do operationally (that is to say, without being controlled). In other words, these variables are included in the measurement situation, but they are not specifically measured. They are allowed to exercise their effect only as part of the mission scenario context; they serve as background for the variables being manipulated. Rather than controlling them, the investigator permits these variables to function in almost a random fashion. However, if these variables are introduced randomly to represent real-world occurrences, it is necessary to repeat the measurement situation (trials) often enough to ensure that the frequency distribution of these variables is not grossly distorted and is in fact representative of what

occurs in the real world. This increases the complexity of the situation of course.

Of course, if one were to ask a question specifically of one of these factors, e.g., what is the effect of high input loads on system efficiency, then it would be necessary to develop a more or less traditional experimental design to accommodate that variable.

If it is impossible to introduce these system factors into the laboratory, then it is necessary to conduct research in the operational environment. However, it is the author's contention that although it may be difficult to include all these factors fully in a laboratory study or even in a system simulation, it is possible to include at least some of them.

Our aim in developing laboratory tasks must therefore be to reproduce the operational environment as much as possible. Within that simulation it is possible to arrange variables in classic experimental designs. System variables serve as the context within which other variables of more immediate interest can be manipulated.

The system-oriented investigator pays a price for his orientation, particularly in the laboratory situation. Fidelity to the operational situation--realism--becomes the criterion of acceptable research. For example, it no longer becomes possible to use "naive" college students recruited directly from the classroom; they must be highly trained as Kidd's subjects (Kidd, 1959) were trained, or they must be operational personnel if operational tasks are used. Almost certainly data collection will require much more time. Team situations will become more common because most systems of any complexity are organized around teams.

With all the difficulties attendant upon incorporating system-relevant factors into controlled experimentation, why should the researcher bother? Because failure to include these factors in the laboratory tends to produce very artificial task situations. The reader may have had the experience of attempting to model a synthetic task after a real world original. In the process of "cutting it down" so that it would fit into a "reasonable" experimental situation, he often simplifies it so it will not require extensive prior training of naive subjects; compresses it so that it will fit within a 30-minute test situation; modifies it so that it can be presented in a group rather than an individual session (thus

again saving precious time); extracts the essence of the real world task so that it becomes only a symbolic analogue of the original rather than a concrete task. Nonetheless, when the study is written for publication the researcher will claim that the results throw light on the way in which the original task is performed. Further comment on this point should be unnecessary.

The effect on system output is the criterion of significance for personnel variables. Since the performance of its personnel is designed to support the output of the system, that output becomes the ultimate evaluational criterion. This means that if a personnel variable does not influence system output significantly, it is unimportant on a system level, however significant it may be on an individual basis. Suppose for example one conducted an experiment in which assembly line production is contrasted under two conditions--with and without piped-in music. Two measures are applied: ratings of job satisfaction (individual performance level); and number of units produced (system output). The difference between the music-no music conditions is highly significant (statistically) in terms of job satisfaction ratings; but the output measure produces only small and variable differences. One would have to conclude that the variable was insignificant from a system standpoint.

One cannot automatically assume that a variable affecting individual human performance will automatically have a corresponding effect on the system output. The effect of individual operator performance on the system output may be reduced by intervening factors which cancel out the effect. This often happens in large, complex systems in which chains of activity must occur before the terminal output is achieved. What might be a significant effect in a single operator system may be insignificant in a multi-operator one. A performance modification in one small link of the chain may be diminished to insignificance by counteracting factors by the time the output is accomplished. It is possible in such a chain-series, the closer to the final link (the system output) that such a change in operator performance occurs, the more likely it is to have a major effect on that output. The more direct the linkage between a human response and the system output, the greater effect that response may have.

One might object that with this philosophy variables of great importance to the individual performer could be overlooked. (It is true that judged on that criterion many human performance variables would drop out.) However, if all

system elements influence each other, if a variable is sufficiently important on an individual basis, it will inevitably produce an effect at the output level. We would not go so far as to say that if a variable is unimportant at the system level, it is also unimportant at the individual level, because this would be manifestly incorrect. Although one should not ignore important individual personnel variables, one should treat them with a certain reserve until their significance to the total system is determined.

Performance must be measured on both individual and system level. Because the system functions on four levels (individual, team, subsystem, system), performance must be measured on all four levels; any description of system performance is incomplete if only individual or only system performance is measured. The traditional behavioral research study is incomplete because it gathers data solely on the individual; the same is true of the usual engineering study because it gathers data solely on equipment.

Another reason for measuring on all these levels is that, as indicated previously, one can only determine the practical significance (i.e., meaningfulness) of a human performance or a personnel variable by evaluating its impact on the system output. The difficulty that arises in actual system measurement is that at the individual/team level one is measuring an output derived from relatively few inputs; at the subsystem/system level one is measuring an output derived from many earlier or concurrent lower order inputs. As one consequence of this, at subsystem/system levels many more human and machine outputs are mixed together, whereas at the individual/team level, although there are machine elements in the operator's performance, the human elements are often more manifest. It therefore becomes more difficult to determine the effect on system output of a particular human response at the operator level.

If we contrast laboratory measurement with measurement in the operational environment, we see because of the number of levels in system operations that:

1. In the usual laboratory research the whole problem of relating human performance to a system output does not exist, because the system--in any of its forms, the operational system, system simulation or system analogue--is not included in the measurement scenario.

2. In the usual laboratory research machine outputs merely aid the presentation of stimuli and the researcher need not be overly concerned with machine characteristics (except in tracking, perceptual or decision making research where the human response is implemented by the machine). Where an operational system is studied, its machine elements as part of the system being measured must be considered in terms of research parameters.

3. In the operational environment performance at the individual level is often separated from the system output level by time and intervening steps with their correlated outputs. If X_1 is the individual output and Y the system output, then

$$Y = (f) X_n \cdot \cdot \cdot \cdot X_4 \cdot X_3 \cdot X_2 \cdot X_1$$

The difficulty in relating X_1 to Y increases as X_n increases. In the absence of a controlled experimental design in which X is systematically varied, the only practical means of analysis is correlation of X_1 with Y; and the more the intervening steps between the two, the lower the correlation is likely to be. On the contrary, in the usual laboratory research the time intervals between stimuli and subject response are usually quite short (see for example the study by Bourne (1957) in which the feedback interval is measured in fractions of a second) and the steps between them few because:

a. The researcher often cannot afford lengthy time intervals and many intervening steps.

b. He knows that if he expands his measurement situation he is much less likely to secure significant differences or high correlations.

By abbreviating his measurement situation he finds it much simpler to demonstrate significant relationships among variables. Unfortunately, because of the abbreviations he produces in his test situation, these relationships often bear little resemblance to operational reality.

The system concept emphasizes measurement in the operational environment.
It does so for several reasons:

1. The operational manned system is the measurement model.

2. It is difficult (although not impossible) to include total, functioning systems in more controlled (e.g., laboratory) measurement environments.

3. Efforts to reproduce system tasks in controlled environments often--although this need not necessarily be so--lead to highly artificial situations which bear little relationship to their operational models.

4. As we shall see later, there are many measurement questions--and not merely those of a research nature--that can be answered only in the operational environment. This last is perhaps the most important of these reasons.

The system concept emphasizes evaluation of performance. It has been pointed out that evaluation is inherent in the system concept because it is necessary to determine whether performance standards are being or can be met. Evaluation tests must be continuously performed during system development (to guide its proper development) and, after the system becomes operational, during its operations (to control and stabilize effective performance). These tests which are of three general types--Exploratory, Resolution, and Verification testing--will be described in Chapter Four. For the moment it is sufficient to say that evaluation testing represents a category of behavioral measurement which is distinctly different from the traditional hypothesis-testing of controlled experimentation. These differences have significant implications for the manner in which system-oriented measurement is conducted and its results analyzed.

Implications for Laboratory Research

If the reader has gathered the impression that system-oriented research must be performed in the operational environment with operational systems, and that in consequence the system concept has no meaning for laboratory research--this impression is incorrect. Logically the most desirable measurement is made in the operational environment, but sometimes this is for a variety of reasons not feasible. It is therefore unrealistic to contemplate a Human Factors research program without laboratory effort. How then should laboratory research be performed so that it fits into the system orientation?

1. The questions asked. The questions asked in this research should focus on the relationship between individual and system parameters. By this

we mean that if the study is, for example, one on the effects on performance of varying delays in providing feedback to subjects, the null hypothesis would be reformulated to read: Does it make any difference to individual and system performance of increasing (or decreasing) feedback delay? In this formulation the researcher does not ignore the individual, but he places him in context with the system. The system context is supplied by organizing experimental tasks into a mission scenario.

2. Equipment requirements. Operational equipment is not required in the laboratory situation; indeed, at one extreme no equipment at all may be necessary, since the principles of system functioning apply in all manned systems and can be studied in manual systems as well as in those that are highly mechanized.

It is perfectly feasible for the investigator to develop an imaginary (synthetic) system which is defined by the following (as a minimum):

- a. A mission scenario consisting of a series of tasks which must be performed over time to achieve a specified system goal.
- b. A system output or product which is required by that goal and which is derived by integrating and transforming outputs from one or more sources (preferably several).

The researcher can develop his system so that one subject's outputs are acted upon by another subject who makes use of those outputs to develop his own. Each such transformation process can be considered a subsystem. One reason for including transforms in the laboratory study is to permit the evaluation of the significance of individual responses in terms of their effect on a higher order output. Transformation can be made to occur by requiring some form of output coding or recoding or by making the first subject's responses one of the bases of a decision made by the second subject.

The researcher may of course run into difficulty in simulating the one-man system in which the transformation occurs solely at the individual level. Transformations in one-man systems must be accomplished with the aid of machine functions which make it difficult to simulate such systems economically. Where

the transformation is accomplished solely on a manual basis, it may be "cleaner" because the researcher does not have to take account of machine interactions with behavioral processes, but it is, however, somewhat unrepresentative of the many transformations that do make use of machines.

The reader may ask how many transformations define a system. The answer is purely a guess on our part. Presumably the more transformations, the more complex the system analogue; but a very simple system can be represented by a single transformation.

Transformations can occur at all system levels (individual, team, subsystem, system). They often occur sequentially, as when information must be gathered from subordinate levels, filtered (interpreted) and integrated by superior levels. Or they may occur concurrently at the same or different levels. Transformations are related to the allocation of functions and superior-inferior levels of authority, but are most likely to occur at points in the system at which different individuals (in a team) or different subsystems interact, i.e., where an output from one function/team/subsystem must interact with another. (Examples of transforms as reported in the literature and of system-oriented measurement in general are described in the Appendix.)

3. Teams. System simulation within the laboratory strongly emphasizes team situations, because the simplest way of accomplishing a transformation is by having different individuals operate upon each other's outputs. In addition, most operational systems of any complexity are multi-operator systems, if only because increasing system complexity imposes too great a burden on a single operator.

In studying individual performance in a system context the researcher is perforce studying team behavior. This is a bonus for behavioral research in general because comparatively few studies have so far been performed using team situations. Most academic research has focussed on the individual, not only because of Psychology's individualistic orientation, but also because creating the team situation is inherently more difficult.

4. Training. Operational systems are exercised by trained crews. The common practice of employing subjects unfamiliar with the tasks they are to perform is completely unacceptable to system-oriented research. Measurement of

naive subject performance merely determines how long it takes him to learn. Often such subjects have not fully learned by the time they must be released. The employment of naive subjects also requires the construction of highly abstracted, simplistic tasks which must be simple if they are not to require prolonged training sessions. Subjects' performance under these conditions cannot be extrapolated to that of operational personnel.

If the task to be performed is an actual operational one, then operational personnel trained in that task must be secured; if a synthetic (artificial) task is employed, subjects must be given extensive training on that task before testing begins. The use of a mission scenario means that the subject will often have to learn several interrelated tasks; and since the tasks will be meaningful ones, the usual practice of giving, say, a fixed number of training trials (e.g., 20) or training to criterion where the criterion is something as absurd as one perfect trial, will not do.

CHAPTER THREE

CHARACTERISTICS OF PERSONNEL SUBSYSTEM MEASUREMENT

Almost all the behavioral research one reads about in the literature deals with controlled experimentation (CE). CE is measurement performed under highly controlled conditions found primarily in the laboratory and emphasizes manipulation of variables as part of hypothesis-testing. Not often described in the readily available literature is a whole important genre of measurement which is largely ignored by experimentalists and statistically oriented psychologists. This genre the author has termed personnel subsystem measurement (PSM).

Although as we saw in Chapter Two, CE can accommodate the system orientation, much of the measurement performed in relation to developing and operational systems involves PSM. When correctly performed, PSM implements the system approach to measurement. PSM is measurement of personnel performing the total task or job (or aspects of these) in the context of or in reference to the actual work (i.e., system) environment.

Although there is no complete dichotomy between CE and PSM, the major point of overlap being a common research function, there are enough significant differences between the two to warrant thinking of PSM as a distinct form of human performance measurement rather than as an applied (i.e., "weak") form of CE. These differences are summarized in Table 1. The following discussion follows the headings in Table 1.

Purpose

PSM has many more purposes than does CE. Where the measurement reference is the system, many more questions arise than when the reference is the individual. The necessary and sufficient purpose of CE is to discover the mechanisms responsible for the performance of the object or event being measured. The application of this knowledge is someone else's responsibility; in PSM that application is inherent in the measurement.

PSM goals are highly pragmatic:

1. To determine the feasibility of an approach.

SUMMARY OF CE AND PSM CHARACTERISTICS

Characteristic	CE	PSM
Purpose	<u>Research:</u> To determine the effect of specified variables	<ol style="list-style-type: none"> 1. Determine feasibility of an approach 2. Select the optimal alternative 3. Determine capability for performance 4. Evaluate effectiveness of performance 5. Solve personnel subsystem problems 6. To perform needed research
Methods	Manipulation of variables	Primarily exercise of the system
Hypotheses	Essential	Largely unnecessary
Performance Criterion	None needed	Both individual and system criteria required
Tasks	Usually synthetic or abstracted functions developed by experimenter	Specified by system; can be modified only with great difficulty
Standard of Proof	Statistical; conformance to theory	Statistical and practical
Unit of Reference	Individual	System
Measures Employed	Individual/Group	Individual/Team; Subsystem/System
Statistical Analysis	Largely testing significance of differences	Largely correlational; some significance testing

2. To select the most effective alternative (e.g., design or procedure).
3. To determine capability to perform.
4. To evaluate system and system element effectiveness.
5. To solve problems arising in the personnel subsystem.
6. To perform needed research.

When the investigator measures in relation to the initial five goals, he is performing what we call "operational measurement." When he conducts research, he is performing "research measurement."

The CE researcher can stop when he has supposedly unearthed these mechanisms, but the PSM investigator (in his operational measurement role) cannot, because PSM's action-orientation demands that where the performance of an object or phenomenon deviates from system requirements, it must be modified to bring it into consonance with those requirements. If the object or performance meets those requirements, the PSM investigator accepts it as it stands; he does not search for causal mechanisms except as part of problem solution. Nevertheless, research is also important to PSM. It is often necessary to learn more about how certain factors affect personnel subsystem performance. When that research is performed, whether in the laboratory or in the operational environment, the methods employed (to the extent that they can be employed) are those of CE.

Methods

To discover explanatory mechanisms, CE must manipulate variables. The experimenter extracts the major influencing variables on the basis of his hypotheses; he manipulates these by arranging his measurement situation. PSM in its operational measurement mode rarely manipulates variables or assigns subjects to contrasting groups because ordinarily there are no treatment conditions to be considered. With one exception. When contrasting conditions are inherent in a system scenario (e.g., the system must function under daylight and nighttime conditions or on sea and land), these conditions will be contrasted.

On the other hand, PSM emphasizes exercise of the complete system scenario (i.e., the mission for which the system was programmed by its developers). Exercise of that mission scenario is a highly complex activity, as exhaustive as the most complex experiment.

Hypotheses

Because the experimenter looks for explanatory mechanisms, he must develop hypotheses about how his variables will function in his measurement situation. He then arranges that situation to test these hypotheses. The PSM investigator has no need to develop hypotheses because for operational purposes PSM is concerned only with whether system personnel perform in accordance with system requirements.

Performance Criterion

Since CE is concerned only with explanatory mechanisms and rarely has a system reference, it does not evaluate the effects that derive from its manipulation of variables. It therefore need not specify in advance that a particular level of subject performance is required (this is not the same as specifying a statistical level of significance); whatever occurs is sufficient. Performance criteria are inherent in operational PSM because that measurement is directed by the question of how effective system performance is or will be.

Tasks

In CE the tasks to be performed by subjects do not ordinarily derive from a specific system although they may occasionally be developed to represent tasks performed in a type of system, e.g., Kidd's air traffic control tasks (Kidd, 1959). Since the experimenter's intent is not to provide information about a specific system, he has great freedom in developing his measurement tasks. Therefore, (and this is something pointed out in Chapter Two) these tasks are often so abstracted that they bear no relationship to operational tasks. Whether or not this is unfortunate depends on whether it is important to the researcher to extrapolate the results of these tasks to the real world. Because there is no necessary relationship between experimental tasks and any particular system, the generalization of data from such tasks to a specific

system is highly limited, but conversely these data may generalize (however haltingly) to systems in general. The generalization is broad, but shallow.

Since PSM in its operational measurement mode is always directed at a particular system, it takes its measurement tasks from that system and they must replicate the actual characteristics of the operational tasks as closely as possible (fidelity). Fidelity is irrelevant for CE. Operational PSM data have maximum applicability to the individual system but less generalizability across systems. PSM generalization is narrow but intensive.

Standard of Proof

Both CE and PSM employ statistical standards of proof; these are sufficient for CE because CE intends only to test whether or not a particular hypothesis is statistically verified. Statistical standards are insufficient for PSM because PSM's action-orientation requires the solution of a system problem or an evaluation on which some action will be based. In addition to the usual statistical techniques and (where applicable) elaborate experimental designs, PSM must be concerned about a pragmatic standard of proof: Is the difference between system requirement and actual performance sufficiently large (however statistically significant) to make a practical difference to the system output?

Unit of Reference

One of the major differences between CE and PSM is the latter's orientation to the system. There is no necessary system orientation in CE and in fact most behavioral experiments are oriented around individuals or groups rather than systems. As a consequence, the application of performance data in CE is usually to the individual or group without reference to the system in which they perform. PSM views personnel performance in the working environment as resulting from attempts to satisfy a hierarchy of system requirements; performance can therefore be understood only in terms of those requirements. (There is of course performance directed solely at satisfying individual goals, e.g., maintenance of bodily functions, and to study this performance CE methods may be wholly satisfactory.)

In PMS the meaningfulness of individual and team performance data lies in its effect on the higher order structure (the subsystem and system) in which performance occurs, and the interplay between that structure and that performance. Consequently, data in PSM must be gathered not only at the individual/team level, but also at the subsystem/system level; and explicitly or implicitly, the effect of the former upon the latter must be examined.

These differences between CE and PSM have major effects upon the way each measures. This does not imply that one is better than or to be preferred over the other. It is simply that differing purposes and conditions of measurement create different measurement situations and that the investigator must apply the methodology appropriate to his situation. In the real world of systems this usually means PSM.

Measures Employed

Because CE studies individuals and groups, measures are taken of both, but primarily of individual performance; group performance is studied much less frequently. PSM measures at the individual and team (work-oriented group) levels, but, in addition, when it is properly conducted, measures are taken at the subsystem and system levels. The latter may appear as less behavioral than the former (e.g., number of targets downed, number of rounds fired) because they summarize not only behavioral but also equipment outputs.

Both CE and PSM make use of quantitative measures, the former being perhaps more molecular and sophisticated than the latter, because it is possible with CE to arrange the measurement situation to make use of such measures. PSM measures are more descriptive because there is much greater emphasis in PSM on normative data, which is inherently descriptive.

Although both CE and PSM employ qualitative data, the latter employs more such data and these play a more central role in PSM than in CE. No controlled experiment should be completed without debriefing subjects, but often even that little is not accomplished. Qualitative methods (interviews, questionnaires, ratings, critical incidents) play a much greater role in PSM because understanding of explanatory mechanisms must in part derive from the cooperation of the test subject. Since it is impossible as in CE to arrange contrasting conditions so

that a clearer understanding of the effects of events and phenomena can be derived, PSM finds it necessary to make use of the experience of these events/phenomena gained by subjects during the measurement situation. In CE research in which the goal is the determination of generalizable principles, subject reactions to the measurement process are not considered particularly important. In PSM, where system personnel form an integral part of the system and can influence the effectiveness of system performance, it is highly desirable to secure their reactions to the measurement situation.

Measurement produces data, and data are what we scrutinize to derive conclusions from that measurement. One might think that all data are the same, and in one sense they are (i.e., reaction time (RT) is the same measure, however the conditions of gathering that reaction time may vary); but data differ in terms of the purposes for which they are gathered and the measurement operations performed to produce them. Most important, they differ in terms of the variables included in the measurement.

All data reflect the selection of particular variables influencing those data. If RT is gathered in a laboratory under highly controlled conditions, the specific RT values are likely to be different than if the researcher gathers RT data under operational conditions (if he were in fact to do so, which is improbable). When experiments are performed to test hypotheses, the data they produce reflect only the experimental conditions included in the hypothesis-test. In consequence, such data are likely to be narrow and less accurate representations of actual operations than data gathered in the operational environment. Some narrowing of the conditions under which CE data are gathered is necessary, because otherwise the data collection task might be impossibly onerous. However, operational data are less constrained by controls than laboratory data, and are consequently less precise or more confounded.

Data may be derived from the following types of measurements: (1) controlled experiments; (2) PSM testing; (3) PSM normative data gathering. In each of these the purpose of the data gathering and the operations performed differ. We have already spoken of CE and PSM testing. The immediate purpose of PSM normative data gathering is to describe as completely as possible the status of an object or phenomenon. The underlying purpose of this data gathering is to provide a data base to the designer/developer of new systems and to assist in the

implementation of some action involving these systems. In normative data gathering, like PSM testing but unlike CE, there is no manipulation of variables. Data are gathered by selecting those measures that appear operationally to be most meaningful. Moreover, because the measurement situation is not arranged, the resultant data reflect all the variables that would ordinarily influence the data.

The investigator will be interested in two types of PSM normative data:

1. Those describing systems, differences among systems, and the relationship between personnel performance and system parameters.
2. Those describing personnel in the performance of their tasks, which usually includes an equipment interface. This last has been termed "human performance reliability" data, although actually these data are focussed as much on accuracy and adequacy as on consistency (see Meister, 1978).

The particular normative data one collects depends on the questions one seeks to answer:

1. System normative data may be collected in response to the following questions:
 - a. What parameters distinguish one system from another?
 - b. Is there a typology of systems and does personnel performance vary with different types of systems? For example, Meister (1977) has suggested a typology based on differences in the amount of indeterminacy present in system operations.
 - c. How does personnel performance vary as a function of the particular system parameters found as a result of answering question 1?

System normative data are very like that produced by traditional hypothesis-testing, i.e., general conclusions buttressed by empirical data, which in this case would probably be based on correlations rather than the usual tests of significance of differences (e.g., Analysis of Variance).

2. Personnel performance normative data may be collected to determine the probability of task/job accomplishment as a function of:

- a. Equipment characteristics.
- b. Job characteristics.
- c. Aptitude for job-related skills.
- d. Personnel skill level.
- e. Experience level.
- f. Motivational level.

In contrast to CE and PSM test data, personnel performance normative data are expressed in the form of tables or nomographs. An example of such tables is provided in Appendix II. Although conclusions or principles can be derived from tabular compilations of such data, they are not specifically expressed in the tables.

Statistical Analysis

CE and PSM are also differentiated by their methods of statistical analysis. In CE variables are tested by arranging contrasting conditions of presentation. As a consequence, the statistics preferred by experimentalists is that testing the significance of differences, the most common technique being analysis of variance in its various formats.

On the contrary, the PSM investigator who measures in the operational situation has little opportunity to arrange contrasting conditions and he often employs correlational analysis (although when he works in a laboratory or can arrange contrasting conditions in the operational environment, or can select contrasting conditions from that environment, he also uses significance of difference statistics).

There is no doubt that significance of difference statistics are a much more powerful tool than correlational analysis, not because this additional power is inherent in the former technique, but because the arrangement of contrasting conditions permits the investigator to zero in on the mechanism or factor hypothesized to cause a given effect. Correlations merely suggest that an association between two or more variables exists; because the investigator is unable to extract the effect of possibly interactive variables, the conclusions he can draw from a correlation must be more tentative than those of the experiment. However, the process of arranging contrasting experimental conditions by eliminating interactive variables tends, as has been indicated previously, to make the experimental situation somewhat simplistic and sometimes artificial.

CHAPTER FOUR

PSM MEASUREMENT QUESTIONS

The purposes of PSM imply certain questions, the answers to which will achieve those purposes. In this chapter we describe those questions and the methods of securing answers to these questions. We have organized these questions by: stage of system development; training program development; system operation; and system maintenance. Research questions particularly pertinent to PSM will also be discussed, but these are obviously not the only research questions relevant to PSM.

Each question will be discussed in terms of the following topics which the investigator should consider before initiating any study:

1. Why is the question pertinent?
2. To what stage of system development or operations is the question most relevant?
3. Does the question require operational or research measurement, and why?
4. Will the study results be generalizable and to what extent?
5. Must variables be manipulated or not; if yes, what are these variables?
6. What part of the total system must be exercised or simulated?
7. Is a performance standard needed?
8. Does the question require a special measurement design or measures?
9. What special problems may arise in answering the questions?
10. What information (in addition to answering the specific question) can one derive from the study?

A. Questions Related to System Development

1. Do personnel possess the capability to perform certain functions to specified levels? Are system design concepts feasible from a personnel performance standpoint?

These questions are asked during the initial phases of system development. With increasing technological sophistication we come closer to developing systems that exceed human capabilities or stress them unduly (in individual functions only, of course). For example, a new system design may require personnel to make perceptual discriminations which are at or about the threshold of perceptual capability. Whether personnel will be able to make these discriminations is unknown. If the behavioral literature does not supply definitive information on this question (and usually it does not), it will be necessary to conduct a study to answer the question. This type of study is termed an Exploratory Test.

Question A1 is peculiar to a specific system design and can be answered only in the context of that design. Regretfully it must be reported that the developer often relies on the operator's mythical capability to overcome severe demands; hence the frequency of Exploratory testing is not as great as perhaps it should be.

Exploratory tests are almost always confined to the Predesign or very early design phase because afterwards the questions are moot; the developer is committed, whether or not personnel can perform the necessary functions at the level required. (It is rare that a function cannot be performed at all.)

Control in the sense of CE is unnecessary in Exploratory testing because there are no contrasting groups. All that is required is reproduction of the (anticipated) system characteristics. Simulation fidelity is necessary because question A1 is specific to a type of design; hence the essential characteristics of that design must be incorporated in the measurement situation. Only enough of the system to permit personnel to perform the system function (if they can) need be simulated, not the total system; no manipulation of variables is needed, nor is there any requirement for comparison with a control group. However, the investigator must compare

personnel performance with the system requirement which is in question because he does not know if it can be accomplished. If an explicit (quantitative) standard or even an implicit one (i.e., the developer's concept of what is acceptable performance) cannot be specified, A1 cannot be answered. Often the information available about standards is insufficient and imprecise because system developers have not thought systematically about what must be done by personnel.

As the measurement situation becomes increasingly specific, the results of the study lose generalizability; hence Exploratory testing does not provide a great deal of usable research data, although it does tell us a bit about what the human is capable of.

On the other hand, since some part of a system configuration must be simulated in order to provide a setting in which subjects can perform, the investigator can secure additional information about the adequacy of that configuration (e.g., from a human engineering standpoint), the particular problems the subject experiences, the nature of his errors and failures--all of which can be fed back into an improved design. In general, every developmental test involving human performance in which the system configuration is realistically simulated provides an opportunity to evaluate not only that performance but also the system configuration used in the measurement situation.

2. Which of two or more alternative system configurations (e.g., equipment designs, manning arrangements, operating procedures, etc.) is more effective from a personnel performance standpoint?

This question may arise both in the Detail Design phase of system Development and in the later Operations phase. If personnel performance is particularly critical to the effective functioning of the system, so that the selection of a design or procedure must be based on that performance; and if two or more designs or procedures are available and the selection cannot be made on empirical or logical bases, then a Resolution test must be performed. Sometimes this question is combined with A1 because there may be some doubt as to whether either alternative will satisfy system requirements.

Although personnel performance is critical to the answer, the apparent reason for the Resolution test may often be an engineering one, (i.e., the configuration as a whole is in doubt rather than the operator's capability to perform).

Since the Resolution test is an operational one, study results may not be very generalizable. The Resolution test resembles CE in the sense that it involves a comparison of configurations; therefore it is necessary to control the conditions under which each alternative is tested. As in Exploratory testing, no manipulation of variables is required.¹ Only those aspects related specifically to the alternative configurations being compared need be simulated or exercised in a Resolution test.

In conducting this test some consideration should be given to operator capability and experience with configurations of the type being tested, because these factors can determine the absolute level of performance achieved. This last is of interest, although the major question is a comparative one; a performance standard, although not crucial for answering A2, is desirable, because, as pointed out before, neither configuration may satisfy system requirements. Because of the need to control test conditions, a study design in which subjects are their own controls (all subjects perform under all conditions and order of presentation is systematically varied) is desirable (although may not be practical). To the extent that system characteristics are simulated faithfully, it is possible to secure information on subject responses to human engineering features of the system, and any difficulties subjects may experience.

3. Does the system (in any of its developmental forms, e.g., drawings, mockups, procedures, prototype or production hardware) satisfy system requirements from a personnel standpoint?

This question arises because every system is developed to satisfy specific requirements (e.g., to fly X miles without refueling; to process N

¹Which is not to say that it is forbidden. Some investigators may do so because somehow it makes the study apparently more respectable (academically, that is). The unnecessary manipulation of variables is not illegal; it is merely inefficient.

amounts of mail each hour); and therefore it is necessary to determine whether the new design has been successful in accomplishing those requirements. These requirements are those of personnel performance which are, however, bound up with and derive from the overall system mission.

The question is answered by the Verification test, whose purpose, like that of the Exploratory test, is specifically evaluation of performance in relation to a requirement (standard). The difference between the two tests is that Exploratory testing involves only a limited number of system functions; the Verification test involves the entire system.

Unlike the Exploratory test, but like the Resolution test, Verification testing can be performed at any stage during system Development and Operations. At earlier developmental stages it is possible to evaluate the design of individual items of equipment, making use of drawings, mockups and prototype equipment (Meister & Rabideau, 1965). Although much of this testing is quite informal, one can consider it as verification testing to the extent that it follows a procedure which is amenable to scrutiny by someone other than the evaluator himself.

At a somewhat later stage of development, when a hardware prototype has been developed, the purpose of the test may be to verify the adequacy of that prototype and to gain information about system feasibility, as the basis of a decision to proceed (or not to proceed) with further engineering development.

Still later, but prior to formal acceptance of the system, the developer may be required to demonstrate the adequacy of his product by means of an Operational System Test (OST); or the system procurer may run his own OST as the basis for system acceptance. The Department of Defense prescribes (Directive 5000.3) a series of developmental and operational tests prior to a system "going operational"; each has its specific purpose.

There are for example, two general types of military test and evaluation (T&E): developmental (DT&E) and operational (OT&E). DT&E is performed by

the system developer and the military development agency. OT&E is conducted by the user and/or by a major field agency, with operational and support personnel of the type and qualifications of those expected to use and maintain the system when deployed. Within OTE&E there is initial operational test and evaluation (IOT&E) performed on pre-production prototypes or pilot production systems; and follow-on T&E (FOT&E) conducted in the field using the production system to verify system performance and operating costs; to validate correction of previously identified deficiencies, and to refine tactical employment doctrine and personnel/training requirements. Occasionally developmental and IOT&E tests are combined.

Thereafter, during the remainder of system life, Verification tests of one type or another (to be described in relation to question C1) may be performed.

The studies performed to answer question A3 are operational ones; they are specific to an individual item of equipment or a system; hence their generalizability to other systems is low. Nevertheless, we hypothesize that data useful to systems in general or to particular types of systems can be gained if results of a number of system verification tests are combined, particularly if a common theoretical framework and measures have been used. One might examine those data from the viewpoint of whether, for example, a design or manning principle common to all systems tested has been particularly successful (or the reverse). To the author's knowledge comparative studies of similar systems under test (e.g., the intercontinental ballistic missiles Atlas, Titan, Minuteman) have never been performed, but there is no reason why they could not be.

The measurement design for a Verification test is comparatively simple, involving only a single condition (unless the mission scenario requires alternative operational modes which must then be compared). Hence no manipulation of variables or conditions is usually necessary. Nevertheless, the care required in exercising even a small system for test purposes is formidable. The test requires exercise of the entire system, in as faithful a reproduction of the operational scenario and conditions as possible;

if the system is exercised in a non-operational mode, the test loses much of its value as a predictor of ultimate operational performance. Reproduction of the operational mode must include personnel factors: system personnel who will exercise the system being tested should either be those who will later run it operationally or have similar characteristics.

Since the purpose of the test is to verify compliance to system requirements, a precondition for the verification test is a set of quantitative standards. For various reasons such as indifference to or ignorance of personnel factors this may be difficult to achieve. Consequently Verification tests are often performed to implicit and hence rather imprecise personnel standards; test sensitivity is thereby reduced and the answers secured are tenuous.

A number of problems may arise in conducting the Verification test. Among the most serious are:

a. Functioning equipment where involved (as it almost always is) may break down, interrupting the measurement process and reducing the opportunity to secure a large enough sample of personnel performance. If the breakdown reveals a serious design fault, major equipment modifications may be necessary; procedures for operating the equipment may have to be revised, so that data gathered on previous performances of system personnel are no longer completely valid.

b. A system under test may be operated in ways that differ from its intended operational deployment. Or not all of its functions may be exercised. This may arise because test personnel wish to experiment with variations in the system configuration. If this happens, the investigator cannot fully answer the initial measurement question as it pertains to operational utility.

c. The personnel exercising the system during the test may not be its intended operational users (e.g., soldiers, factory workers). The value

of some tests has been lost by employing the engineers who developed the system to exercise it; since these are usually more highly qualified (and certainly more knowledgeable about the system) than the anticipated using personnel, an incorrect prediction of user performance will be generated.

Because the IOT&E and FOT&E (in contrast to their earlier development or test versions) deal with the entire system and all its functions, it is possible to secure considerable information about many facets of that system. Previous Verification tests (see Askren and Newton, 1969) have examined the following:

- a. Human design considerations relative to operability and maintainability;
- b. Adequacy of technical publications used by personnel to operate/maintain the system;
- c. The work environment or other conditions that affect personnel performance.
- d. Adequacy of the manning estimated as needed to utilize the system;
- e. Adequacy of training received by test personnel.

Deficiencies in any of these areas can be noted; investigations of causal factors for these deficiencies can be initiated; prospective solutions to these problems can be tried out and validated.

4. How do system developers develop systems? How do they make use of Human Factors inputs? What is the relationship between system characteristics and operator performance?

These questions are pertinent because they go to the very heart of the

Human Factors goal of assisting in system development. To do so it is necessary to provide system developers/designers² with the behavioral information they need to make more adequate design/development decisions. This in turn requires the behavioral specialist to have an intimate knowledge of how the designer functions. This knowledge is necessary because the designer is usually the sole authority with regard to anything relating to his design (or at least has the most strident voice with regard to accepting or rejecting recommendations of design characteristics). The specialist needs first to determine what behavioral information the designer needs; then, what information he will and can utilize; next the format in which it should be presented; and finally how he utilizes that information.

In order to make meaningful recommendations for the incorporation of behavioral principles in design, the specialist must ascertain the relationship between individual system characteristics (e.g., various equipment arrangements) and the operator performance resulting from those characteristics. We still know very little about this relationship.

Question A4 calls for research rather than operational studies, because answers do not apply to a single system or equipment. If one knows for example that a particular equipment characteristic or configuration always requires a particular kind of maintenance activity, this information applies "across the board." For that reason the answers to A4 should be highly generalizable.

Since we are dealing here with research rather than evaluation, the manipulation of variables is not only pertinent but necessary. In the studies

²By system developer we mean one who is responsible for the planning, approving and managing of system development; he may be the head of the customer's project team, the Chief Engineer or head of an engineering group. By designer is meant those who carry out the developer's plans at a detail level by drawing designs for equipment; planning spares requirements, performing the engineering calculations and writing the operating procedures. The informational needs of developers and designers probably differ, the latter requiring more molecular information than the former. For convenience we shall henceforth refer to the "designer" only, with the understanding that this phase also includes the developer.

performed by Meister et al. (1968, 1969a, 1969b, 1971) and Askren and Lintz (1975) one can see some of the variables around which this type of research can be designed. These include (a) individual differences within the designer population (e.g., the system level designer vs. the equipment level designer); (b) the type of human factors input (e.g., equipment-performance relationships, skill level, availability data, etc.); (c) the format in which human factors inputs are presented (e.g., verbal, graphic, quantitative, etc.); (d) the amount of behavioral information provided and its sequencing in development.

Since the research is not system-specific, it is unnecessary to attempt to replicate the hardware characteristics of any particular system. Rather the researcher is attempting to investigate a development process, and the measurement environment selected for such an investigation should be that of an actual engineering department or a situation which simulates the essential characteristics of the design process. For example, in one study (Meister et al., 1968) missile ground equipment designers were hired to design (on paper only, of course) a ground fuel pressurization system to the experimenter's specification. The design effort was performed in the designer's own office which was of course part of the company's engineering department. Since most of the inputs and outputs of the design process are symbolic (cognitive), in the form of engineering drawings (although mockups too can be developed), a special physical environment is not necessary.

Because these are research questions and moreover deal with operator performance only indirectly, performance standards are irrelevant. (Performance standards for design adequacy are very gross.) Nor does the resulting research require a special experimental design specific to the questions asked; any of the usual experimental designs involving experimental and control conditions can be employed; which one is selected depends on the variables at issue.

Although the equipment problems to be solved need to be based on an actual operational system, one aspect of the measurement situation does require a high degree of operational fidelity. In simulating the design process it is necessary to reproduce the characteristics of typical inputs

to that process and to require customary design outputs from engineers. For example, it is necessary to develop a realistic system requirement document and to ask for certain types of analyses and drawings which one would ordinarily receive in response to such a requirement (e.g., schematics, flow diagrams, tradeoff analyses). In studying design many methods can be used to secure data, including observation of actual system development activities, interviews with designers, paper and pencil tests employing written/graphic design problems, ratings of the adequacy of design outputs, etc. More controlled methods may make use of simulation as well as experimentation both in the laboratory and in the actual system development environment (i.e., the engineering department).

The simulation of anything as complex as the design process presents difficulties. There are two critical elements affecting the design response: differences among designers; and the nature of design tasks. The first is important because design is a peculiarly individual (i.e., covert) process, despite the use of project teams on major projects. We are unable to do more than make gross differentiations among types of designers, (e.g., system and bench-level engineers); and such a dichotomy probably covers up critical individual differences among them (see Hughes Aircraft Company, 1978). The second is important because design requires a wide spectrum of functions. Since the design process is usually a lengthy one (for systems of any complexity) the experimenter attempts to simulate the process (and speed it up) by abstracting what he considers to be the critical elements of that process. But the abstraction may eliminate certain other essential features of design that have not yet been identified.

This is an especially fertile area for fundamental PSM research; it has been hardly touched (the few studies performed and the conclusions reached require more detailed treatment than is possible here); but its potential utility cannot be overestimated. Unfortunately the importance of the question is not too well recognized by those supporting behavioral research.

B. Questions Related to Training

1. Has the necessary training been accomplished? Is the training adequate?

Question B1 does not mean, has a specified course of study been given? But rather, have trainees learned appropriate³ skills on the basis of that training? It is the training program that is being evaluated and not any individual student's proficiency, although obviously whether or not individuals (in aggregate) have learned (as demonstrated in performance) is the basis for determining whether or not the program is effective. Whether or not individual X has learned is a pertinent question but only to individual X.

Question B1 arises as soon as training is initiated. The very first training for a truly new system is usually factory training, i.e., training provided to potential users by the system developer at the factory or test range. This training is apt to be somewhat less systematic and formal than that later established by the using agency (the customer) on its own (basing that training of course on the factory predecessor). Factory training often precedes the OST which serves as the acceptance test for the system since using personnel who will conduct that test must be trained. However, B1 can and is often asked throughout system life, because systems become updated, curricula are refined, student population characteristics may change (with greater or lesser aptitude than previous inputs); all of these make it necessary periodically to determine whether training still satisfies objectives.

Question B1 is obviously not peculiar to manned systems but may be asked in any educational context; for example, it has been asked increasingly about the adequacy of the American primary and secondary educational system and whether that system is doing what it is supposed to do. In the case of public education as contrasted to technical (e.g., military or commercial/industrial) training, the problem of training evaluation is considerably exacerbated because educators, politicians and the general public have ideological viewpoints and rarely agree on what the system requirements are or should be. This is much less true for technical training.

³The reference is always the system. That training is appropriate which satisfies system performance requirements.

There are additional differences between training for a new military or commercial/industrial system and public education. In the former, if the system is genuinely novel, there is little backlog of experience and knowledge on which to base decisions such as: how long training should be; what training methods should be employed; what should the sequence of training be. (Major military system changes are infrequent; however, less comprehensive changes are much more frequent.) Consequently these decisions are likely to be based on the "cut and try" principle. Public education on the contrary has changed relatively slowly (despite the propagandizing of new approaches) and there is a longer "track record" to use as the basis for these decisions.

Naturally question B1 applies not only to a total curriculum but also to its individual segments.

The evaluation of a specific training program can be answered only by operational study because it is peculiar not only to a specific set of tasks and skills demanded by the system but also to trainees with a particular set of capabilities and experience. The answers provided are not generalizable, therefore, except in the general sense of confirming that a certain amount of training given in a particular way does (or does not) lead to skill acquisition. (Few comparative studies of actual training programs are performed, however.) One can also ask whether a type of training environment, such as a simulator, or a training medium such as programmed instruction, trains effectively. When the question is asked in this general sense, it becomes a research problem (see question B3), the answers to which are highly generalizable.

As long as the evaluation is an operational study, no manipulation of variables is required in the evaluation of an individual training program. (The situation changes however when one is contrasting different training methods or media within a single training program; here the training program serves merely as a context for the experimental variables which are treated as in question B3. If on the other hand one is contrasting two distinctly different training programs, the study become a Resolution test.)

The design for evaluation of an individual training program may be handled in several ways:

a. One may ask how much improvement trainees exhibit as a result of training over their initial (pre-training) state. In this situation trainees are tested before, during and following training; if the differences between post-test and pre-test are significant, it suggests that an amount of something has been learned.

The defect of this kind of design is that no mention has been made of any performance standard which trainees must satisfy. Lacking a comparison with that standard, all one can say is that some learning has been achieved, but whether trainees have learned what they should and to a prescribed criterion of adequacy is quite unknown.

b. Alternatively, trainees are tested only once, at the conclusion of training (and/or at the conclusion of individual curriculum segments); and performance is compared with some standard of proficiency (which may be a quantitative score, an expert judgment of capability--transformed into a quantitative rating--or accomplishment of some tasks representative of the job). This design is a very common one, but the difficulty is that since an objective standard is usually lacking one can again surmise only that some training has been accomplished but not the amount of that training.

Control groups which receive no training and serve as comparison conditions are not usually employed in operational evaluations, because first, there is no pool of subjects available to serve as a control, since all personnel entering the training pipeline are there to learn; and secondly control groups are unnecessary since specialized training will always produce performance significantly greater than would be produced by a group not receiving training. (The concept of a control group in evaluation of training is appropriate only when there are alternative ways of developing skills than by means of a specified curriculum.)

If what is being trained are skills which can be demonstrated only by performance, it will be necessary to exercise the system, subsystem or equipment (whichever is relevant to the skills to be demonstrated). A surrogate- simulator- of the operational equipment can be utilized in place of the actual equipment if the simulation is sufficiently faithful. Only if knowledges (e.g., background information, theoretical concepts) are to be tested⁴ will exercise of the physical system be unnecessary; in this case the system can be represented symbolically (i.e., verbally, by drawings or mockups).

Answers to question B1, if it is asked of the factory training course developed before OST, may be provided (partially) by the OST. If one assumes OST to be the operational environment to which factory training should be responsive, deficiencies in training can be inferred from difficulties manifested by test personnel who were trained in the factory course, and by interviews with them to solicit their opinions concerning the adequacy of their training. Judgments of training adequacy made in this way are inferential only and should be considered tentative if only because they may be confounded with the effects of hardware problems occurring during OST. For example, if design inadequacies are found that negate procedures learned by personnel, the training may be downgraded when the adequacy of the hardware is at fault.

Questions of training program adequacy must also be asked once the curriculum is established. No curriculum is entirely stable and as data are fed back from the operational system (concerning the adequacy of personnel sent from schools to the operational system) progressive refinements should be made in the curriculum to bring it more closely into accord with operational requirements.

Each major modification should be evaluated. This can be done in either of two ways: (1) either the modification can be specifically and individually tested against the previous method (Resolution testing); or (2) the efficiency

⁴However, in contrast to non-technical public education, it is most unusual in system training for knowledges alone to be required of the trainee; he almost always has to demonstrate in performance of some task the application of any knowledge he has acquired.

of the total training program (with the new modification incorporated) can be evaluated and compared against its previous efficiency (a form of Verification testing). In (1) a formal test is conducted with the previous method and the new one as comparison conditions. In (2) the modified curriculum is tested against data describing the performance of the previous training program. The ideal situation is to follow both procedures, but this is rarely done; if any evaluation at all is performed of a modified curriculum, evaluators usually follow the second procedure (to the extent that they have data reflecting the previous program) and infer from the overall status of the curriculum (i.e., the ultimate success or failure of its students, together with instructor judgments) whether or not the modification has been effective.

The effectiveness of a training system implies that what is taught is actually relevant to operational tasks. This is of course a critical question but one which is not usually addressed in measurements of training efficiency. Relevance is usually assumed or (less frequently) is evaluated on the basis of informal judgments. We consider this question to be related to the concept of transfer of training and it will be considered later in that context.

In order to evaluate a new training program properly

a. Personnel performance standards related to operational tasks are needed, from which training measures can be developed. Many so called standards (at least those used in military systems) are, in a performance sense, not standards at all; they describe the knowledges needed for performance, but not the performance itself. Moreover, the standards referred to describe individual trainee performance only; there are no quantitative performance standards that can be applied to the training program as a whole.

b. A measurement setting is needed which realistically reproduces the demands of performing the job operationally. The evaluator cannot ordinarily exercise the actual system fully in a training mode in order to provide these conditions; the question therefore arises of how much less than complete system fidelity he can accept and still provide conditions reasonably representative of the operational situation.

If the criterion is performance on the operational job the training environment as a setting for training evaluation inherently lacks some degree of validity. It follows therefore that training system efficiency measured by student performance in the training environment should always be followed by testing the graduate input to the operational system. However, this procedure creates additional problems (to be explored later) and is rarely implemented.

Training program adequacy obviously cannot be evaluated in a binary way: yes, the program is efficient; no, it is not. All training programs are almost certainly deficient in some respects which a sufficiently sensitive measurement methodology would reveal. Lacking that, the measurement should at least indicate the difficulties trainees have and the functions in which they are most and least proficient.

2. Does performance transfer from the training environment (usually the school) to the operational job? How does that performance transfer?

The first question demands an operational study, specific to a particular system, training program, types of personnel and mode of training. The second requires a research study focussed on determining the principles by means of which transfer can be maximized. In this discussion we shall focus on the first question, not because the second one is unimportant, but because there are already excellent treatments of the topic.

What do we mean by the phrase "transfer of training"? Traditionally transfer has been defined as the application of skills learned in one task or in one context to learning of a second task (Hovland, 1951). Other definitions are possible however. In the one which is of particular interest to PSM, transfer refers to the extent to which the trainee can apply his skills once learned to the operational job for which he was trained. There is no contradiction between the customary definition and the one employed here because almost always some additional learning of that job in the operational environment is necessary (which suggests that no training program can be completely efficient; if it were, the graduate would perform with maximum

effectiveness as soon as he entered the job environment). What is transferred is a skill learned in one environment (usually a school) to another environment, the operational one; the demands of the latter are usually greater than those of the former. From the standpoint of evaluating a training program the essential question is, can the trainee perform where it counts (however well he performs when tested in the school?). Studies performed in the military indicate that many personnel are inadequately prepared to function in the operational environment (e.g., Steinemann et al., 1968).

If it is true that transfer depends on stimulus and response similarities between the training and the operational environment, inadequate transfer may result from the fact that the school does not provide stimulus-response connections sufficiently similar to those of the operational system (a problem of fidelity). Or (less likely) the school may not have identified the skills which should be taught. If either of these hypotheses is correct, personnel trained in the job environment should perform more effectively on that job than those receiving an equivalent amount of training in the school environment. A test of this hypothesis would be to train one group in a set of skills at school and compare its performance with that of another group (equated in capability) which learned the same skills for the same length of time on the job. From an evaluational standpoint, however, it is almost impossible to equate formal school time with informal (or even formal) on-the-job observational and practice opportunities. For example, the same instructors do not train both groups; indeed, the OJT group may have no instructors (in a formal sense) at all.

To determine whether learned skills transfer from the school to the operational environment it is necessary to

a. Compare the performance of school graduates in Operations with that of an aptitude-equated control group which has received an equal amount (obviously not type) of training on the job. This means measuring two groups in two environments, the school and the operational system.

b. Measure the performance of both groups in the operational environment, either by setting up special evaluation tests (almost never possible) or by measuring normal routine activities. In the latter case chance variations between performance opportunities available to the different sets of subjects may confound results.

In actual practice the effectiveness of training as it transfers to the operational job is almost always measured (when it is measured) in a non-performance manner; for example, measuring the capabilities of operational personnel by using paper and pencil knowledge tests (which assumes that knowledge is equivalent to performance); or by the ubiquitous supervisory ratings. The results of such tests are then compared with some standard (usually quite imprecise) of what performance should be. Training adequacy is inferred as a function of the discrepancy between the standard and the measured value.

Whatever the method used to evaluate operational performance, the results of these evaluations often reflect dissatisfaction with personnel performance. From the many complaints of performance inadequacy by supervisors (at least in the military⁵) it would be reasonable to infer that much training does not transfer; and then of course the question to be answered is why, a question that is answered by investigation and not by formal test.

In summary, the general practice is that performance of operational systems is measured or (more usually) inferred and compared with an expectation of what that performance should be; then, if there is a sizeable discrepancy, as there often is, the training program is blamed. However, outside of experimental research studies, no evaluation of training systems is very systematic and based upon sound methodology. The exception is the evaluation of the effectiveness of simulators (most often aircraft simulators); and these are devices that implement a training program rather than the program

⁵ Whether such complaints are equally prevalent in non-military systems is not known. In commercial/industrial systems inadequate personnel can be discharged; and many non-military systems (e.g., construction, transportation) assume no responsibility for training their personnel. In military and social-benefit (e.g., Civil Service) systems it is much more difficult to get rid of incompetents.

itself. Formal measurements to determine how much has been transferred from the school to the job are rarely made, because of the difficulties already cited. Arrangements have been made in the military to provide feedback from the operational system to the schools to suggest where training deficiencies exist, but again this feedback is based largely on informal impressions and has not generally been very valuable in upgrading training quality.

Transfer in the traditional sense is also involved in the establishment of a curriculum. Ideally the curriculum developer arranges the sequencing of tasks to be learned so that there is maximum transfer between part tasks (which come earlier in the training sequence) and the larger jobs into which the part tasks fit. However, this sequencing is almost never based on previous transfer measurement (or even on principles derived from transfer research); nor is there any attempt to measure the degree of transfer involved between part and whole tasks in actual training programs.

Transfer can also be used as a means of comparing the efficiency of alternative training modes, e.g., the increasingly common use of simulators to replace operational settings as the training context. Aircraft simulators have been used to replace many hours of actual flight training (NAVTRAEQUIPCEN, 1972). Weapon system simulators (e.g., sonar trainers) make it unnecessary to exercise actual systems frequently and they also permit practice of emergency procedures that would be very hazardous in actual operations. What is transferred is again a skill developed in a particular training environment (e.g., the simulator) to performance in the operational environment. In this context question B2 implies the following three sub-questions:

- a. Is a particular training environment (e.g., the simulator) effective?
- b. How much training in that environment will replace how much training in the operational environment?
- c. Are training environments of a general type (e.g., aircraft simulators) effective?

The first of these sub-questions is an operational one; the third, a research question; the second can be answered either as an operational or a research question or subsumed under (a) because any comparison of training environments involves some amount of training given in each environment.

The study design to answer these questions follows the paradigm of traditional transfer of training designs. For example, the adequacy of an aircraft simulator can be determined using as shown in Table 2.

Table 2. Study Design for Transfer of Training Studies

<u>Group</u>	<u>Training Environment</u>	<u>Training Hours</u>	<u>Performance Environment</u>
Experimental	Simulator	N	Air
Control	Air	N	Air

The performance (criterion) environment is the air, because that is where operational flight occurs. The experimental group receives N hours of simulator training; the control group receives the same number of hours learning and practicing the same tasks in the air. If the experimental group performs as well in the criterion environment as does the control group, one would be justified in saying that the simulator was effective and that N simulator hours have a training value equivalent to N hours of flight training.

Variations are of course possible: one might compare N simulator hours against N+3 hours in the air, in which case if the simulator were effective, it would have more than equal training value compared to the operational environment. Or the experimental group might receive N hours in the simulator and P hours in the air, compared with a control group receiving N+P hours in the air. Or one might compare a number of experimental alternatives (e.g., N experimental groups) against a control.

This is of course only a very simple treatment of a very complex topic. However if one asks only the specific questions: is the simulator effective,

and to what extent, this treatment of the topic may be sufficient because one is not asking about explanatory (i.e., theoretical) mechanisms for which a more complex research design is necessary.

These comparisons imply of course training on the same task content to the same performance standard, although obviously students in the different settings will not be trained in precisely the same manner. Whether training in one mode or environment is identical with that given in another is irrelevant, as long as a specified criterion performance is demonstrated in the operational job.

The criterion of transfer is always performance in a relevant operational job/environment. In some cases it may be difficult to measure criterion performance operationally because the operational system is not available to the investigator or does not permit sufficient control. In those cases an intermediate environment (i.e., one similar to the operational one) may be used. For example, a sonar simulator might be used to substitute for an operational sonar to evaluate the effectiveness of sonar technician training (Mackie, 1978). Such ersatz situations do not however provide a completely satisfactory answer to the questions raised in this section, because they all differ from the operational environment to some extent.

In operational studies of transfer only the training environment and the amount of training given in that environment need be varied. In research studies of transfer other variables can be manipulated simultaneously. If, for example, the investigator wished to determine the interaction between simulator training mode (e.g., variable task sequences) and amount of transfer, he might subdivide his experimental group into two or more subgroups, depending on how many task sequences he wished to explore. Such sophistication is not ordinarily found in operational transfer studies, however, because it is unnecessary.

Subjects for transfer of training studies should ideally be equated on variables that affect their learning proficiency (e.g., intelligence, aptitude). This involves a degree of control that is ordinarily difficult to achieve, if the subject population is small.

3. How do the effects of training in one mode or with one medium compare with those of another mode or medium?

Examples of training mode are: classroom or programmed instruction;
of medium are: film or texts.⁶

This is a research question for two reasons: (a) training specialists do not usually have the time, inclination or money during their development of a training program to answer this question; (b) the answers sought are intended to be generalizable beyond the immediate system. The question arises because the information is needed if the developer is to construct the most efficient curriculum (i.e., training personnel to required performance in the shortest time period and at the least cost) for a particular purpose. To do this the developer must know whether programmed instruction for example will train students to the same proficiency level in a shorter time than conventional classroom instruction; if so, it makes sense to select programmed instruction as the training method to be used. Unfortunately there is no well established set of principles or data that permits the training specialist to select rationally among his alternatives and in consequence he makes his selection largely by feel. Hence the need for research to develop these principles/data.

We touched upon this question briefly in discussing transfer, because the transfer paradigm is often used to study the question; the reader will recall as an example the simulator vs. flight training. The trouble one runs into in implementing this paradigm is that ideally all subject personnel should receive the same training under both sets of conditions, thus avoiding the necessity for equating subjects in different treatments; but since the same stimulus inputs must be learned under both training conditions, learning with the first medium or mode presented makes the second essentially irrelevant.

⁶There is no generally accepted taxonomy of training modes/media, but obviously the following (by no means exhaustive list) must be included: classroom instruction, simulators, self-paced instruction, on-the-job training, etc. as modes; computer assisted instruction, programmed texts, audio-visual devices, motion picture film and video tape as media. It is of course possible to subcategorize each of the above.

Two groups must therefore be used, each receiving the same training with a different mode/medium; but this means that the two groups must be equated on variables that affect training efficiency. With a large enough population this difficulty is not overly constraining.

One could of course ignore the transfer factor completely and simply train two equated groups of subjects with two media/modes to the same training criterion (e.g., an accuracy score). Presumably that group which learns more quickly has utilized a more efficient medium/mode. Such a comparison, while easier to implement, is deficient, however, because there is no guarantee that either group will perform satisfactorily in the operational environment or that the group which performed less well in training (and whose medium/mode might therefore be considered less efficient) might not perform more effectively in the operational environment.

Since what the student learns must be related to some system context, the specific nature of what is being trained reduces the generalizability of this type of research and makes it necessary to repeat media/mode comparisons with several types of training material to ensure the validity and generalizability of resulting conclusions.

Because the training materials in such studies describe aspects of the operating system, it is necessary to represent those aspects in the research. This means employing a hardware surrogate of the system (e.g., simulator, mockup) as part of those training materials. Often however the representation is purely symbolic (e.g., photographs, diagrams), particularly when knowledges alone are being taught. Even with knowledges evaluation of training adequacy requires some sort of performance demonstration. In measuring criterion performance the actual system or a simulator (but only that unit relevant to the skill being measured) must be exercised.

Since the transfer research design involves a comparison, a performance standard is not absolutely required, but since a standard presumably exists, the investigator will obviously be interested in determining which mode/medium

produces performance closer to that standard. It is necessary also to ensure that what is learned (regardless of mode/medium utilized) is relevant to some operational job or jobs in general; if it is not, the study results will be essentially irrelevant also. For example, it is meaningless to compare two training modes in terms of ability to learn nonsense syllables, unless nonsense syllables can be shown to be required in some operational performance.

As in other studies of transfer, the criterion performance measures should ideally be collected in the operational environment. Because it is difficult to follow school graduates to and to secure access to them in System Operations, this is not often done, and it is much more common for researchers to use a surrogate such as a simulator or the actual system exercised in a quasi-operational manner, for example, flight performance on a special test range. These substitutes, while closely paralleling the operational environment, are not quite the same. Because of this it is impossible to say with complete certainty that mode/medium A is better than mode/medium B, although if the criterion performance test has been carefully programmed the investigator can have a reasonable confidence in his results.

The trouble with training studies is that criterion performance is influenced by so many variables: the nature of the training materials and the trainers; the amount and type of training given; its relevance to system operations; where the performance is measured, etc. Consequently any single study may present only a partial answer and that study must be validated with other training materials, subjects, training methods, etc. before one can reasonably develop a general principle.

4. How faithfully must the training environment reproduce the operational one?

This is a question which for obvious reasons permeates all training; a decision on this point is made every time a training program is developed,

but almost always without substantive data as the basis for the decision. In cost terms alone it would be desirable to reduce operational fidelity to a minimum, because complete fidelity (as seen in some elaborate simulators), is horribly expensive. Question B4 leads to research rather than to an operational study because the training program developer does not have resources or time to pursue an answer specific to his program. That answer has most generally been sought in connection with fidelity of simulation, an entire literature on which has been created. Nevertheless, a definitive answer to the fidelity problem even as regards simulation alone has never been established.

The reason for the question is obvious: maximum fidelity seems logical, if one wishes maximum transfer, although some studies have shown that for certain training situations maximum fidelity is unnecessary (Grimsley, 1969; Prophet and Boyd, 1970). The nature of what is trained appears critical to the amount of fidelity required. Because of this the extent to which one can generalize fidelity relationships is limited.

This is a question in which the analysis of variables prior to conducting the study is most important. Because of the large number of variables that may impact training efficiency, it is critical to determine (or at least to hypothesize) those stimulus-response dimensions for which fidelity is most likely to be important. This permits the investigator to limit his research design to what can reasonably be studied.

Because question B4 relates to the operational environment, its study design is basically a transfer design. The researcher compares groups trained in environments or with materials of differing fidelity and tests them in the operational (criterion) environment to determine which group performs more effectively. The transfer studies described previously are also inherently fidelity studies because the differences between the operational and training environment are differences in fidelity. As in other transfer studies performance in the operational setting is the criterion; whatever dimensions and degrees of fidelity are associated with more effective performance in that environment are assumed to be more effective for training.

Since what is trained derives from some system, it is necessary during training to exercise that system (or its surrogate) but usually only for those elements of interest to the study (e.g., displays); certain aspects of the system may be exercised only symbolically. This may create difficulties when measuring criterion performance because in actual operations those elements that are being specifically tested are embedded in other elements and cannot be extracted in operational performance without unduly distorting the system.

If the fidelity study is performed in the context of actual system training, it is possible to uncover information about those system aspects that are especially difficult for students; with which aspects they are most likely to make errors and the nature of those errors. This information may be used to refine the training program around which the fidelity study has been conducted.

C. Questions Related to System Operations

1. How well do system personnel perform relative to requirements?

The results of OST determined that the system (including personnel) met its performance requirements. This was the basis on which the system was accepted by the customer. It is necessary, however, to continue evaluating the system and its personnel throughout system life because both may change over time and continuing operation; it is important, therefore, to determine on a continuing or a periodic basis whether their performance still meets original requirements. If it does not, some remedial action will be needed. (So far it has been implied that only the system changed, not the requirements. It is, however, possible for requirements to change even after the system has been developed and for a system, therefore, to require re-verification.)

Tests to answer question C1 are to be considered Verification tests conducted after the system becomes operational. There are three types of such Verification tests:

a. Continuous system evaluation, the essential characteristic of which is that measurements are taken continuously of normal, routine operations. For example, recording component scrap rate on a continuous basis. Some drivers habitually record the gasoline consumption of their automobiles.

b. Periodic system evaluation, which makes use of the same routine operations noted in (1) but samples these on some basis (e.g., time, subsystem or even personnel). Particularly important subsystems may be monitored more frequently than less critical ones. An example would be to collect scrap records sampling particular assembly lines, components, personnel or time periods (e.g., 6-12 shift).

In continuous or periodic evaluations the system is not asked to perform any exercises specially designated for measurement purposes. In

c. Special system evaluations, the system performs special exercises which, although they may not differ from routine operations, are specifically for evaluation purposes. For example, when an automobile is driven on a test track to determine its gasoline consumption, the vehicle is driven more or less as it is ordinarily, but the goal of the run is measurement. Other examples of special system evaluations are military exercises (e.g., war game) and the OST.

The three types of system evaluation are not equally desirable. Continuous evaluations are preferable; if these are not feasible because of cost, for example, periodic evaluations are an acceptable substitute. Least desirable is the special evaluation, because it is "one-shot" test and may be outmoded by system changes and events. It may, therefore, be less valid than the other two.

C1 is, of course, an operational question and the answer to the specific question has no research implications (although see the last section of this chapter). In order to answer the question the system must be exercised, of course, but except for the special evaluation no non-routine operations must be performed. All the investigator need do in continuous and periodic evaluations is to establish the scope of the subject population (the system, or particular subsystems to be measured) and the personnel performance measures he will apply in the measurement.

Unfortunately, when data on system performance are routinely gathered under the control of personnel who are not behavioral specialists, the measures taken are often output measures only, describing subsystem and system output and ignoring personnel performance. Unless behavioral specialists are part of the measurement team, it is unlikely that behavioral data will be gathered even when (as is often the case) personnel performance is required to conduct the test.

An example is the collection of maintenance data in the Navy using the 3M system, the focus of which is on description of the failure rather than on the performance of the technician who corrects the failure (Williams and Malone, 1978). In production systems scrap rate is gathered rather than data on errors made by production workers. Admittedly it is much easier to collect hardware data than personnel data because hardware outputs derive more or less automatically from system operations whereas personnel actions require a deliberate effort at data collection.

Because the purpose of the tests is simply to compare performance against a requirement, (although if they have been performed previously, it is also possible to compare present against previous performance, so that one can derive trend data), no manipulation of variables is necessary. No sophisticated study design need, therefore, be established (although a statistical design for data analysis is a prerequisite). Since, except for the special system evaluation (and even then only rarely) system operations are not modified, no selection of personnel or their assignment to treatment conditions is necessary (or usually possible). The entire system is, of course, exercised but the particular aspects of those system operations to be selected for measurement (e.g., a particular subsystem, functions, etc.) depends on the investigator's interests. Obviously these evaluations cannot be performed without precisely described quantitative standards. It is possible to gather data without such standards, but the end result is not an evaluation but simply the gathering of descriptive normative data.

Some of the difficulties that confront the PSM investigator in undertaking system operations were noted earlier and need not be repeated here. The major problem in interpreting the results is that of relating behavioral measures such as task accomplishment and task response time to system outputs which are

non-behavioral. The difficulty can be related to the distinction which can be made between process and product measures where product is the output of process. Process measures can be roughly equated with behavioral measures; product measures, with system outputs. By the chronology of the situation, behavioral measures precede output measures. Behavioral measures describe tasks; many tasks may have to be performed to achieve a single output. How then do we correlate a multiplicity of behavioral measures with a single (or relatively few) system outputs? (Statistically this may present no problems, but the logic of the relationships found may be unclear even when statistically identified.) If there were only a single task, one could discern a consistent relationship with the output; but with multiple tasks, each of which may have several measures, the interpretation of the correlation is confusing at best.

2. Is the system ready to perform as required?

Some systems, especially (but not exclusively) military ones, do not exercise all their functions on a continuing basis. Certain functions like emergency ones are performed only rarely and routine functions only occasionally. For example, Naval ships engage in cruises only periodically. In the civilian sector, agencies dedicated to emergency functions (e.g., civil disaster, hospital emergency rooms) fortunately have the opportunity to exercise those functions only rarely.

Such functions are difficult to evaluate on a continuous basis. It is necessary, therefore, to determine whether the system is ready to perform those functions when called upon. For example, prior to the time they start a cruise Naval vessels are required to demonstrate their operational readiness.

Tests to answer question C2 have much in common with system verification testing, because one must determine whether the system can perform in accordance with requirements. The difference between Operational Readiness Testing (ORT) and other verification tests like OST is that the investigator may be required to conduct his measurements without exercising the system operationally. The results must then be extrapolated to the operational (untested) situation.

There are two ways of performing ORT:

a. One can attempt to simulate the functions that cannot be operationally exercised. For example, Civil Defense calls disaster drills in which simulated victims are bandaged and rushed to emergency rooms. Most weapon system mission simulators permit exercise of combat and emergency functions that cannot actually be performed without hazard to participants (e.g., the introduction of simulated malfunctions). The military performs combat exercises in which their units "play" at combat. It is possible to carry the war game simulation even further by abstracting the purely intellectual (decision making) features of combat functions and performing these symbolically. An example are the computer-assisted war games played at military institutes. In a manual form these have even entered civilian life in the form of games in which one "re-plays" historic battles.

b. Another alternative is to measure the performance of functions related to the non-performed functions and to extrapolate the results to the latter.

Suppose, for example, one wished to evaluate the operational readiness of an air surveillance system for combat; one might assume that system performance in routinely detecting and tracking fly-overs of friendly aircraft would be representative of the way in which it would perform in combat.

Or in the hospital emergency room one might measure performance as a function of peak load (e.g., Saturday night victims) and extrapolate to potential larger disasters.

The assumption of a relationship between routine (customary) and non-routine (e.g., emergency, combat) activities is very risky, however, and this procedure, if followed, should be used with great caution, because there may be major differences between responses to routine and non-routine activities. For example, the Navy assesses the operational readiness of its ships partly on the basis that the correct number of personnel with the required ratings for running the ship are available.

Evaluation for the determination of operational readiness involves a prediction based on the extrapolation of data (to a certain extent this is true of all verification tests, particularly OST, but it is especially marked in ORT). What is being predicted is the probability that the system (including its personnel) will perform in accordance with system requirements at some future time. That time period is of some consequence; it is easier to predict immediately future performance rather than events to occur much later. Prediction implies that some validation of the prediction will be made (i.e., that system performance will be measured at a later time to verify the prediction's accuracy), but this validation is rarely performed.

Question C2 obviously requires an operational study, but there is a research relationship also, since there is a great need to develop a more effective methodology for answering the question. In particular a model is needed which describes the elements entering into the prediction and the way they interrelate.

As an operational study whose data are applicable to a particular system, ORT results are not very generalizable, although the model referred to in the previous paragraph, if validated, would be highly generalizable. As in other system verification tests no variable manipulation is required, although the model would include certain variables which could be exercised under different conditions. For example, to develop a prediction of an Army unit's performance it is necessary to consider geography as one determinant of its mission scenario. Alternative operational modes must also be considered: for example, a naval vessel may have, depending on circumstances, the responsibility of interdicting commerce, acting as convoy protection or as a strike force.

Although ORT involves prediction, it must be solidly based on empirical data gathered in the present. This is one of the major failings of ORT in the military. Many of their operational readiness measures are highly subjective (e.g., ratings) and unsystematically gathered, although based on observed performance. Because military raters are not well trained to make evaluations, the ratings may not be very valid. Moreover, the necessity for putting the best light on the evaluation results may bias the rater to produce falsely positive conclusions.

3. How can a problem arising from system verification be solved?

If personnel performance has been shown by a Verification test to be inadequate (i.e., the results do not satisfy system requirements), the cause of the inadequacy must be determined, so that a solution can be found. Since performance deficiencies are often found as a result of testing (although system managers may feel that most of these do not render the system inoperable), this phase of measurement is quite common.

The reader may feel that the investigation of a problem is not measurement, since in many cases it does not involve formal testing or quantitative measures. In our viewpoint, however, it is a special form of measurement because the effort to discover the cause of a deficiency is a deliberate one; because it utilizes special measuring instruments (see below); and because it does produce data, although the data may be produced informally and in non-quantitative form. A more important reason is that action-oriented measurement (which PSM definitely is) is not complete if the system has inadequacies, the causes of which are not yet determined.

Of course, not every inadequacy receives--or needs to receive--the same degree of investigative attention. The first step following data analysis which suggests an inadequacy is determination of its criticality (to system functioning/output); many inadequacies deserve short shrift or minor attention because of their relative lack of importance (see Peters and Hall, 1963). For example, the mislabeling of a control may be reported as a discrepancy, but is usually handled by simply relabeling the control. The assessment of criticality is usually analytic based on the consequences to the system if the inadequacy is not remedied. Those of greater consequence stimulate investigative measurement.

The problem is to set a standard as to what constitutes a critical problem. Here we deal with biases--those of the behavioral specialist who is likely to consider all inadequacies as deserving of maximum investigative effort; those of engineering or system managers who are likely to feel that the human can overcome all problems (and, therefore, nothing need be done). The manager's bias is likely to be more potent in a conflict of this sort.

Investigative measurement is operational and very specific, hence not generalizable, nor intended to be. It obviously requires no manipulation of

variables, nor need the system be exercised, except where a major system problem has been unearthed. By its very nature this type of investigation is quite informal. Performance standards are implicit in the investigation; it was failure to achieve standard that initiated the investigation; but no action regarding the standard is needed unless the investigator suggests that the standard is too stringent. Nor is measurement design required, except possibly a strategy for conducting the investigation. The measurement methods utilized are of great importance, however, for two reasons: (a) they must be tailored to a very specific question and (b) investigative difficulties are much greater than those found in any of the previously described tests. Such methods may include any of the following:

a. Examination of test data, records, operating procedures and technical manuals to determine exactly what happened.

b. Interviews with the personnel whose inadequate performance caused the investigation and/or with their supervisors. (This method is by far the most common.)

c. Questionnaires and rating scales given to test personnel.

d. Special tests of knowledge/skill (almost always in paper and pencil form).

e. Inspection of the equipment (particularly its human engineering characteristics) involved in the test performance.

f. If possible, reproduction by the test personnel of their previously inadequate performance. (This is often not possible if it requires re-exercising a major part of the system.)

This last is most desirable. In some major systems (e.g., the AEGIS computer-controlled missile system) a formal investigation may be conducted in special laboratory facilities in which the variables potentially responsible for the inadequate performance can be manipulated. Unfortunately, for financial, time or logistical reasons, this alternative is rarely invoked.

In most investigations the emphasis is not on hardware but on the individuals involved in the performance; the effort is to elicit from their statements why their performance deviated from that required. The process suggests a form of criminal investigation which, like the investigation of a crime, becomes highly molecular; the focus of attention may be a single operator or a single performance.

Often the development of the appropriate investigative strategy and methods is more complex than the measurement that led to the investigation. It is much more difficult to determine why an event or phenomenon exists than simply to record its existence. (We do not consider theory-making here; problem solution goes beyond the construction of hypotheses since it must lead to some remedial action.) The specific questions asked must be specially tailored to the problem as it appears from the initial test data. For example, interview questions must be phrased with great subtlety to elicit the clues that will reveal the cause of the inadequacy. Investigations to discover the cause of anything must by their nature be indirect, which increases the difficulty of solving the problem. Because test personnel are sensitive to any criticism of their performance, any investigation which implies inadequacy must tread a very narrow line.

It is unlikely that the cause of an inadequacy in an operational system can be discovered by controlled experimentation in a laboratory environment, since the locus of the actions that led to the inadequacy is in the system exercise. (However, a potential solution may be tried out in a simulator before being introduced operationally.) Moreover, the time required to set up a controlled experimental situation is often prohibitively lengthy and the pressure to eliminate the difficulty is very pressing.

Problem solution measurement has the greatest degree of uncertainty associated with it. In other measurements one can be certain of securing some data, whatever the quality of those data; in causal investigations, however, there is no guarantee that the investigator will wind up with anything except egg on his face at the conclusion of the investigation.

4. How does a new system configuration compare with the old?

As has been pointed out, the system is not necessarily static during its 5, 10, or 20 year life. New methods or new equipment may be introduced to modernize it or to improve its efficiency. From this may arise the need to test the effectiveness of the new equipment or procedure. In a sense, at this point the system reverts to its original untried developmental state (at least with regard to the new equipment or procedure) and must be reverified. Consequently the system must perform a Resolution/Verification test--only this time in terms of a comparison between the new equipment/procedure and the old. The aim is to demonstrate that the new equipment/procedure not only satisfies system requirements but works better than the old.

All Resolution testing is comparative. There are, however, several ways system comparisons can be made: (1) a formal test of the new and old configurations; (2) the new configuration only is tested and the comparison is with the old configuration's past performance, represented either by historical data or the impression system personnel have of how well the old configuration performed. Procedure (2) may be performed either formally or informally. The latter is most common; the new configuration is simply adopted into system operations; but even here a comparison is implicit, because the new will not be retained unless it is at least as effective as the old.

Question C4 is purely operational and as such has little potential for generalizing to systems other than the one concerned (although again see the subsequent section of this chapter). The study must be performed with the actual configuration (or a surrogate, such as a highly realistic simulator). Like other operational studies there is no need to manipulate variables, although if both configurations are tested there are two treatments. If a formal comparison test is performed, it is necessary to ensure that the conditions under which the new equipment/procedure is exercised are the same as those with which the old was utilized. Otherwise test results will be spuriously slanted one way or another. This should pose no real difficulty if the mission and environment in which the new configuration is to be utilized have not changed significantly. Among the conditions to be examined is the nature of system personnel under test; if personnel have changed drastically since the previous configuration was operated, the comparison may be invalid.

The formal test design, like that of other Resolution tests, is essentially a comparative one. Unless other variables are inherent in the manner in which the equipment/procedure will be used, no more sophisticated design need be attempted. Because the test is a comparative one, theoretically no performance standard is necessary, but the new configuration must meet system requirements. Where formal personnel/system requirements do not exist, data on the performance of the previous configuration may serve as an implicit standard. In the test of the Launch Vehicle Assault craft referred to in a previous chapter, performance of the new prototype was calculated relative to performance of its predecessor craft, LVTP-7.

As in all operational tests, it is possible to secure information fortuitously, such as the difficulties personnel experience in utilizing the new equipment or procedure (if these difficulties are substantial, they will obviously affect the basic test results negatively); conditions (situations) that may negatively affect the efficiency of the new configuration; estimates of the skill demands levied on the personnel by the new configuration; etc.

D. Questions Relative to Maintenance

1. How do technicians perform diagnostic maintenance?

All equipments, no matter how well designed, ultimately fail; most equipments fail repeatedly during their service life. The availability of the system (defined in terms of the equation $1 - \frac{\text{system downtime}}{\text{mission time}}$) depends on how quickly a failed equipment can be restored to operating status. Performance degradation resulting from operating difficulties (i.e., the operator's inability to operate the equipment efficiently) rarely stops the system from functioning, although it may reduce its efficiency (e.g., range, accuracy, etc.). In contrast, failure to restore a malfunctioning equipment will often shut the system down. From that standpoint, there is more utility if the behavioral researcher concentrates on maintenance than if he focusses on operations.

The emphasis in this question is on diagnostic maintenance, i.e., the strategies and actions required of personnel to resolve equipment malfunctions. This is not to say that preventive maintenance (PM) is not important, but in comparison with the problems associated with remove/replace/repair activities,

PM bulks rather small. The synonyms for PM (routine, scheduled) suggest why this is so: if one can schedule PM, one can specify most if not all the actions required; one can find the most effective actions and more readily train personnel to perform them.

In contrast, because equipment stoppages are difficult to anticipate and when they occur may result from any of a bewildering array of component failures, it is not easy to specify the actions needed to localize the precise component responsible for the failure. Diagnostic maintenance (troubleshooting) is largely a cognitive process, although other capabilities (e.g., the ability to visually discriminate worn or poorly fabricated parts, to visualize mentally the spatial relationships among components, to follow complex instructions and to perform with manual dexterity) undoubtedly play their part. Troubleshooting is, however, above all else a deductive and inductive process; deductive in the sense of inferring from a set of symptoms or status conditions what has gone wrong; inductive in the sense of the technician's applying his previous diagnostic experience to develop a maintenance strategy.

Indeed, it is because the diagnostic procedures performed are in many cases idiosyncratic, because some technicians are much more successful than others in corrective maintenance, that question D1 arises; if one can describe the processes developed by successful troubleshooters, one might be able to teach others to use the same procedures.

Since maintenance is performed during system operations, answers to D1 are most pertinent to that phase of system life; but if one learned how successful technicians perform maintenance, it should be possible to design into equipment those features they make use of. Answers to D1, therefore, have great generalizability, not only to design but as much--if not more--to the training given maintenance personnel.

Research on diagnostic maintenance processes involves three types of variables and hence the manipulation of those variables. These are: equipment characteristics (e.g., the number of components, their hierarchical relationships and interrelationships--how a fault in one component influences the state of others--the availability of built-in test equipment (BITE) or test points); technician characteristics (e.g., intelligence and aptitude, the amount and type of

diagnostic training and experience the technician has, the strategy he typically employs); and external factors (e.g., the availability of job performance aids such as troubleshooting tables, computer aids, the adequacy of technical data such as manuals, failure probability charts, etc., tools and specialized test equipment). At one time or another each of these has been varied in a research mode.

Although the effect of downtime on system capability is massive, the technician working to restore a failed equipment usually confines his attention to that equipment; he does not ordinarily exercise the full system of which the failed equipment is only a part, except to test the functioning of the equipment, once it has been apparently repaired. In research on troubleshooting processes, the level at which the research is directed is almost always the equipment or a module of that equipment; occasionally circuits. In contrast to equipment and system operation which is heavily team-oriented, corrective maintenance is largely individual-oriented. This does not mean that teams of maintenance personnel are not required during corrective maintenance activity (e.g., one man to read a computer diagnostic printout, while the other removes and replaces components); but since the troubleshooting process is largely cognitive it is likely to be concentrated in one (presumably the most proficient) man. Research on troubleshooting processes is, therefore, directed mostly at individuals. (Why then are system-oriented investigators so concerned with corrective maintenance? Because its impact on the system and the effect of the system on it are so great.)

Research on diagnostic processes may be either normative or experimental. Early research was normative: to determine what the maintenance technician did, and thus to find what variables should be investigated experimentally. Thus, in the initial studies performed by Grings et al. (1953) personnel rode Navy ships and observed what technicians did; they asked technicians to keep diaries or to fill out questionnaires. More structured research has involved controlled situations in which faults were inserted into equipment and technicians were asked to find these; or troubleshooting characteristics were abstracted and put into the form of paper-and-pencil tests (e.g., the Tab Test, Damrin and Saupe, 1954); or simulators were developed in which the troubleshooting process could be reproduced without requiring operational equipment (Bryan et al., 1954).

Because the troubleshooting process is highly idiosyncratic, it is difficult to incorporate a performance standard in research on that process. In actual maintenance operations there is, of course, a performance standard which is usually defined as mean-time-to-repair (MTTR). However, MTTR is only an average and somewhat faster or slower performance are acceptable. MTTR is, however, not generally utilized in troubleshooting research because the emphasis is on process rather than output.

Corrective maintenance performance has usually been measured by (1) the time taken to find the malfunctioning component; (2) the number of diagnostic checks made (which obviously influences time); (3) the sequence in which checks are made (whether this is systematic, based on failure probabilities, circuit logic, the effort involved in checking, etc.; or random); (4) ultimate success (whether the fault is finally found or the technician admits defeat. The experimental designs used in controlled studies of diagnostic maintenance do not differ in any significant way from that used with other research topics. For example, two groups of subjects equated in terms of some relevant background variable (e.g., experience) may be given two different types of pre-maintenance training and then tested on a standard set of problems; or the problems presented may differ based on variations in some set of dimensions considered possibly important to troubleshooting success.

Two major difficulties have been encountered in performing troubleshooting research: lack of success in concretely conceptualizing the dimensions and factors involved in troubleshooting; and problems of simulating the diagnostic processes under controlled conditions. Since the process is largely covert, it is difficult to elicit from technicians the factors responsible for their behavior. If the process were not so complex, this difficulty might not bulk so large.

Various attempts have been made to simulate the process. At one time actual equipments were brought into the laboratory and specific faults were then deliberately inserted which subjects were then asked to find. However, it was difficult to control fault insertions; it was found that the insertion of a failed component so often produced a correlated fault in another component that the stimulus condition the subject was responding to was unclear. Maintenance simulators have been developed (Bryan et al., op cit.) but all of them seem to leave out some critical elements of the corrective maintenance situation.

Likewise, efforts have been made to abstract the essential features of the diagnostic process, but in actual practice the process is so complex that analogues of the process are again suspect as not including all relevant features.

2. How efficient is diagnostic maintenance?

The system has maintenance as well as operating requirements; the former are phrased as MTTR's (referred to previously). In order to maintain a specified level of system availability it is necessary that some percentage (often 90 percent) of subsystem maintenance be performed within the specified MTTR. The system's ability to meet this requirement depends almost exclusively on the technician's capability; time to restore the equipment to operational status is affected by non-diagnostic factors, i.e., administrative downtime (e.g., time to secure replacement components; time to transport technicians to the equipment to be maintained; time to write reports) but the impact of these non-diagnostic factors is relatively slight compared to that determined by the technician's capability.

In the same way that any operating requirement must be verified, so must the system's ability to meet specified MTTR's. MTTR data are usually collected during major developmental tests during which equipments fail; and a requirement for verifying MTTR will be a major element in the OST. During routine system operations repair time may also be routinely collected (e.g., the Air Force's 66-1 data collection method) to determine whether the system is retaining its ability to perform to maintenance requirements. The emphasis in operational maintenance data is on restore time; operational personnel are uninterested in the diagnostic process as long as MTTR requirements are satisfied; they become interested in the process only when the requirement is failed.

The determination that a MTTR requirement in a particular system is or is not being satisfied is not usually generalizable; however, a long or short MTTR may be associated with a particular type of system design (e.g., BITE), but this information is very gross.

In the operational study of maintenance (i.e., the verification through demonstration that MTTR requirements are satisfied) variables are not manipulated.

Moreover, the troubleshooting function is among the few operational system functions for which no special arrangements need be made in order to collect appropriate data (provided sufficient time for collecting those data is provided; if too short a time is allotted, it may be necessary to schedule special tests in which faults are inserted and must be found). Since troubleshooting is purely a contingency activity, it is necessary only that the system be exercised in an operational manner and that the necessary instructions/forms for reporting be provided in order to collect the desired data. The responsibility for recording and transmitting maintenance data during routine operations belongs to the technician; this creates certain difficulties which will be discussed later. During Verification tests like OST maintenance data may be recorded by special observers, but the data collection process cannot be allowed to interfere with the process. Since the troubleshooting process is so covert, the observer may experience difficulties in securing more sensitive data than MTTR.

As indicated previously, the technician works only with those elements of the system that are affected by the malfunction. At the apparently successful conclusion of the troubleshooting a system checkout is performed to verify the correctness of the maintenance process, but this will usually involve only the next higher level of the subsystem in which the failure has occurred.

Time is the primary measure employed and the MTTR requirement is the only performance standard utilized. In special research tests an observer may report the sequence of diagnostic checks made and any difficulties the technician encounters.

Some of the problems encountered in attempting to determine whether technicians are satisfying MTTR standards have already been suggested. When a maintenance data collection system is left to the discretion of the individual technician (as is the case with 3M and 66-1) the latter may "fudge" his figures in order to create a better or worse impression than reality suggests. (Better to look good or avoid difficulties with a supervisor; worse if he needs replacement components and the basis on which these are issued is a failed component.) Restore times may be arbitrarily increased to cover non-system related activities (like drinking coffee) or else reduced to ensure that MTTR's are met. Some data may be missing because technicians are reluctant to fill out the required forms (Williams and Malone, op cit.).

The researcher who wishes to make use of operational records to test hypotheses about troubleshooting may find that the special data he is interested in are not being reported by standard reporting mechanisms. The nature of symptoms is usually not reported fully; the sequence of checks is not reported at all (because the reporting system does not require this information). The actual cause of the failure (as distinct from the name of the failed component) may not be reported, either because the technician does not know it, or dislikes reporting in such detail.

Because of the factors reported in the previous paragraph, it is difficult to say what additional information (beyond a suspect restore time) can be secured from a standard reporting method. Under ideal circumstances (these can be found only in highly controlled situations) the following may be determined: (1) restore time; (2) description of failure symptoms; (3) failed component; (4) failure cause. Practically speaking, only the first item can be secured with any certainty. If the investigator utilizes observation and interview methods to secure the data he wishes, all four preceding data items can be secured, as well as the sequence of checks made, the sources of difficulty experienced by technicians, the impact of administrative downtime, etc. But such a personal data collection methodology makes almost impossible time/effort demands on the investigator.

Research Implications of PSM Tests

In this section we will discuss the research data that can be provided by Exploratory, Resolution and Verification tests. These are tests designed to provide answers to specific developmental and operational questions and their data are highly specific. We need not deal with those PSM studies that are specifically research-oriented, since these should (if performed correctly) provide generalizable data.

Why should anyone be concerned about the research consequences of operational PSM tests, since their specificity would appear to reduce their research utility? These tests have, however, one thing laboratory or simulation studies cannot provide, that is, they permit us to see how operational systems function under the actual operating conditions. Since we have little data on and know less about how operational systems function, these PSM tests afford an unparalleled--

literally--opportunity to collect such data. To make such observations requires the application of naturalistic data gathering methods (which are primarily observational) to the manned system in somewhat the same way that such methods are used in animal observation in comparative psychology and ethology. The application of naturalistic methods to the analysis of systems can be much more effective than in comparative psychology, however, since the observer can make use also of available system records and interviews with system personnel.

Before exploring the kind of data we can secure from such tests, we should clarify the usual objection to research data collection in the field, i.e., lack of control. Control permitting the collection of research data does exist in the operational environment, but it is not control in the CE sense, which is accomplished by arrangement of contrasting groups. Control in the operational system is exercised by ensuring conformance to the mission scenario which implies operational equipment, operational personnel and operational procedures. As long as the system functions in conformance to that scenario, the human responses and system outputs it produces (which are recorded as data) are inherently valid. This control cannot be used to test hypotheses but can be used to assure validity. In naturalistic studies hypotheses can be tested (to some extent) by arranging the data produced by these studies.

The kinds of research questions that can be answered from operational PSM tests include the following, which are only exemplary:

1. System Performance

The general question is how does the manned system function in the operational environment? More specifically:

- a. What variables appear to be exercised in the operational system?
- b. How effectively do the systems perform their missions?
- c. How effectively do system personnel perform their functions; how do they contribute to the system output?
- d. What external/internal inputs tend to have greatest effect on system functioning?

e. What inputs/events (if any) tend to interfere with and degrade mission completion?

f. To what extent do systems ordinarily deviate from their specified mission scenario?

g. What are the major observable inputs to and events of system functioning and how do these interrelate with each other and the system output?

h. What problems arise (other than equipment malfunction) that result in incomplete performance of functions or poor quality outputs?

These questions can be asked about the system as a whole or about individual functions or subsystems, e.g., equipment operation; equipment maintenance; logistics, etc.

2. Equipment Maintenance

With regard to maintenance functions, the following questions can be asked:

- a. How successful is troubleshooting generally?
- b. What strategies do maintenance men generally employ in troubleshooting?
- c. What behavioral problems generally arise in troubleshooting?
- d. How does maintenance quality affect system outputs?

3. Training

One can also observe the training program, but here the focus of interest would be in the transfer process, for example:

- a. How much gets transferred from school to the operational job?
- b. Which skills seem to transfer most readily/least readily?

c. When transfer has failed, what seems to have caused the failure?

Answers to training questions are inherently more difficult to secure than the previous ones, because transfer questions require observation of two environments, the training environment and the operational one.

Naturalistic data can be most readily gathered in Verification tests, both special tests and those involving routine system operations, because such data require performance of the entire mission scenario. In special tests (e.g., OST) the degree of control may be somewhat greater than in routine operations because the focus of attention is directed specifically at data gathering. However, special tests, no matter how faithful to the operational model, are probably somewhat artificial simply because they are special (i.e., non-normal).

Some data bearing on system development can be secured from Exploratory and Resolution tests, but the fact that these tests deal primarily with segments or individual functions of the total system and its mission reduces their value. It may be more productive to treat the entire system development process as a system in its own right and to ask of it the same performance questions (item 1 above) but this time focussing on Human Factors efforts as a subsystem of the system development system.

It may appear because of the specificity of the data secured from the various PSM tests that generalizable conclusions cannot be achieved. Although we deal in each case with specific systems functioning in specific ways, there are presumably commonalities among systems or characteristics of systems (e.g., similar structures, information processing channels and processes, etc.) that can be teased out and about which pertinent data can be secured. To discern such commonalities requires some sort of theoretical framework because these commonalities are abstractions. However, even in the absence of such a framework data can be gathered that will perhaps help to develop the framework. Framework and data are mutually interactive and supportive.

The author wishes to suggest that although operational PSM tests are designed to answer specific questions, their results can be used for research purposes, provided that a conceptual framework for collecting and analyzing their data has been developed. Along with principles that can be generalized to systems

as whole, PEM naturalistic data can be used to develop a normative data bank. A data bank describing what? That depends on the questions one wishes to ask, e.g., the frequency of errors of a particular type; the type and frequency of personnel interactions; etc.

It is probably a cliché to say that the first task of a science is to examine naturalistically the phenomena which are its subject matter. Since we say that the manned system is the subject matter of Human Factors, it is incumbent upon us to examine that system as it functions in its normal operating environment.

CHAPTER FIVE

SYSTEM THEORY

In order to perform useful research on manned systems one must have a theoretical structure that deals specifically with such systems. Such a framework should include:

- (1) Ways of differentiating types of systems, i.e., a taxonomy.
- (2) Description of system parameters, i.e., system elements and characteristics.
- (3) Hypotheses about how system parameters interact with each other to affect the system output.

Such a framework might be used to: (a) describe and classify systems in terms of personnel/system relationships; (b) derive hypotheses about the effects of these relationships which can then be looked for in operational systems and/or studied experimentally; (c) compare existing systems with those proposed for new development.

No such theoretical structure at the system level exists. At more molecular levels we have theories that deal with individual functions such as signal detection (Tanner and Swets, (1953)) and vigilance (see Buckner and McGrath (1963)). These are however not system theories (explaining how personnel function in relation to the system) but theories about personnel as individuals.

Miller (1978) has attempted to develop a comprehensive theoretical structure about living organisms as systems and suggests many hypotheses that could be used as a starting point for the construction of hypotheses specifically tied to manned systems. Miller's framework encompasses individuals, groups and organizations but unfortunately not the manned system.

In this chapter we examine some very tentative concepts about how manned systems function.

The first step is to taxonomize the various types of manned systems.

System Types

A major distinction is between military and non-military systems. No less important however is the distinction between the commercial/industrial and social-benefit systems that make up the non-military category. The classification of system types is shown in Table 3. All are manned systems, some being what we customarily think of as man-machine systems (e.g., fighters, tanks), others what are termed organizations, e.g., armies, industries. The system types vary also in their degree of mechanization, the social-benefit systems, for example, being somewhat less hardware-dependent and hardware-sophisticated than the others.

Table 3. Types of Systems

Military	Commercial/Industrial	Social-Benefit
1. Military organizations (e.g., armies, navies, air divisions)	1. System/product developers (e.g., automobile manufacturers, soap producers)	1. Governmental control systems (e.g., legislative, police, welfare)
2. Weapons systems (e.g., fighters, tanks, destroyers)	2. Support industries (e.g., mining, steel mills)	2. Consumer service utilities, (e.g., power plants, water distribution systems)
3. Support subsystems (e.g., medical, logistics, maintenance, transport, etc.)	3. Consumer service industries (e.g., car rental, air lines)	
	4. Individual commercial systems (e.g., automobile, truck, ship)	

Notwithstanding the distinctions made in Table 3, the identities among the system types are striking. On a general level, all possess the essential features of manned systems. (1) All have a goal: to exert force upon an enemy; to design or sell an automobile; to provide welfare or electric power. (2) All are designed (more or less deliberately) to satisfy requirements to which personnel must conform. (3) All proceed through a process of development, although that development may vary in complexity; the more novel the system, the more elaborate its development. (4) All possess an organizational structure involving some sort of hierarchy of system levels, i.e., operator/team, subsystem/system.

(5) All have functions that are differentiated and specialized. (6) All (including social-benefit systems) have a primary component which accomplishes the system goal and a component which supports it by supplying tools, logistics, means of production, computer services, etc. (7) Commercial/industrial facilities develop the hardware (e.g., automobiles, trucks, computers) used not only within their own sector but also by military and social-benefit systems.

Especially noteworthy is the fact that although the primary purpose of individual systems varies widely, each contains many identical subsystems (primarily for support). Medical, communication, computation, transportation, maintenance subsystems are often identical across all three types of systems. No one for example would expect the character of medicine practiced in military, industrial, and municipal hospitals to differ significantly, although there are obvious speciality differences depending on whether the physician is dealing with combat wounds or geriatric diseases.

At the individual personnel level, moreover, the behavioral functions performed by personnel in the different systems are often much the same despite differences in system mission and equipment characteristics. Perceptual discrimination, for example, demands very similar efforts of the operator whether he is scanning a radar or sonar display or scanning a moving stock market quotation display.

Because of these system similarities, it is logical that PSM measurement principles can be applied to all systems. The degree to which formal testing occurs varies, of course; it is certainly more frequent in military (or at least reported more openly) than in commercial/industrial and social-benefit systems. These differences may, however, arise from the different traditions in which these systems develop. Military systems may require more testing because their requirements are often more stringent than those of the others, because they are more technology-intensive than social-benefit systems, and because evaluation is mandated into regulations governing their development and operation. Since social-benefit systems are so closely associated with governmental (i.e., political) organizations which often function in what appears to be an illogical manner, the tradition of testing such systems is

less firmly established, although that too is changing with the proliferation of social programs (e.g., Social Security, Headstart, various welfare schemes) whose evaluation is demanded by taxpayers. The point to be emphasized nevertheless is that the PSM principles derived in large part from military testing situations can be applied to the other types of systems. (The reader interested specifically in evaluation of social-benefit systems might consult Scrivening and Guttentag, 1975.)

On the other hand, there are significant differences within a system category. The most obvious difference is that of size. An organization like an Army utilizes thousands of personnel; the smallest system, a single operator. Because of this the mechanisms involved in the former are perhaps qualitatively different from those required by the single operator system: a management structure which is irrelevant in, for example, the single-seat fighter. Management structures cut across systems of different types and respond primarily to the size differential among systems.

System Clients

Another significant difference between military and non-military systems is that the latter provide a benefit to clients who often have the freedom to accept or reject the outputs of these system, and whose interaction with the system often affects the way in which those systems must be developed, operated, and tested. Military systems have no clients, although with the introduction of all-volunteer forces, even this distinction is breaking down; volunteers (and particularly their families) can in some degree be considered as clients. And of course some military support subsystems (e.g., medicine, recreation) inherently possess clients.

Commercial/industrial systems are more or less competitive (the larger being less so than the smaller); clients may or may not buy a particular brand of car, house, or soap. Military systems compete with each other and with the civilian sector for manpower (and their retention in the service) but beyond this they have no competition; Social-Benefit systems have comparatively little (e.g., one may refrain from drinking the output of the municipal water system but this requires recourse to wells or bottled water).

The implications of the client for measurement is that for non-military systems one must include as factors to be measured (1) the way in which clients interact with system operations and, to the extent that this interaction is important, (2) the desires, needs, and performances of clients as constraints on system performance must be tested. An example is the Social Security Administration, among whose clients are the elderly and disabled (Old Age and Survivor's Assistance) who because of their age and handicaps have special needs that must be catered to by the Administration.

Client factors must therefore be considered in measurement in two ways:

1. Client desires and satisfaction must be considered as criteria of system effectiveness where clients have a choice of accepting or rejecting system outputs. (This applies in part also to Social-Benefit clients who are at the mercy of the system and who must accept system outputs, e.g., welfare recipients, prisoners, energy consumers. Prevailing western social philosophy urges us to take into account their desires and satisfactions; but obviously these are less urgently considered than when the client is completely free to reject the system output.) The investigator must include client's needs/desires as part of the system requirements he uses to derive evaluation performance standards, if the satisfaction of these needs/desires is critical to accomplishment of the system's overall goals.

Of course, since the client has many needs/desires, the system developer/manager may select among these the ones he considers most relevant to the system's overall goals. For example, Detroit pays much more attention to styling than to ease of maintenance considerations, because it has decided it can sell more cars (its primary goal) by satisfying the former than the latter.

2. If the client is required to perform in some manner to implement the system goal, his performance must be measured. For example, in the Social Security Administration (SSA) and Internal Revenue Service (IRS) situations, clients must report various items of information; if they are unable to do so correctly, SSA and IRS processing of that information becomes more difficult and expensive of manpower, time, and money. To the extent that systems have only limited control over client performance (e.g., their availability for

testing) it becomes more difficult to measure that performance; moreover, client measurement may require use of special techniques such as attitude-eliciting devices.

In contrast to military systems, therefore, which need be concerned primarily about system personnel (those who run the system), non-military systems must measure both system personnel and clients. Sometimes these two sets of personnel interact (as in consumer-service industries), so that performance of one can be measured as a function of the other. When they perform independently, however, the measurement burden is increased.

In non-military systems two aspects must be measured: (1) The system functioning to accomplish its primary goal (e.g., fabrication of products to be utilized, production of electric power); and (2) the system functioning in relation to its clients (e.g., selling and servicing the products, distribution of power to clients). Because the military system has no strong client interest, it should be somewhat easier to measure personnel performance in military than non-military systems.

Systems Characteristics

In this section we will endeavor to list the various ways in which systems can be described (see Table 4). These characteristics are enduring qualities that cut across the distinctions made previously among system types. The following list is probably not exhaustive; and the individual characteristics are almost certainly not equally important in terms of affecting system output.

1. Types/number of functions performed (e.g., surveillance, transport).

The purpose the system performs is the traditional way in which systems have been classified. However, system classification by function is not very useful and should be used only as an initial sort. The type of function obviously determines the nature of the system output, but not its efficiency, accuracy, etc.

The number of functions performed also varies among systems. Probably number of functions is somewhat correlated with system complexity and system

TABLE 4

POSSIBLE EFFECTS OF SYSTEM CHARACTERISTICS ON THE SYSTEM

SYSTEM CHARACTERISTICS	SYSTEM EFFECTS
1. Types/number of functions performed.	1. Type determines nature of output but not quality. Number increases system complexity.
2. Number of operational modes.	2. Correlated with system complexity but may improve output.
3. Number of subsystems.	3. Increases communications.
4. System organization.	4. Affects potentiality for breakdown.
5. Number/organization of operator positions.	5. Effects unclear.
6. Number/type/locus of transforms.	6. The more of these, the more likely output will be affected.
7. Number/organization of communications channels.	7. Significant effect on output quality.
8. Output requirements.	8. Affects system feedback and evaluation.
9. Characteristic inputs.	9. Effects unclear.
10. System reactivity.	10. Effects unclear.
11. Degree of mechanization.	11. Effects unclear.
12. System feedback.	12. Increased output quality.
13. System indeterminacy.	13. Significant impact on system processes and output.

complexity may have many consequences, including increased probability of error, output degradation, and system breakdown. However, these last effects are mediated by many factors, of which number of functions is only one.

2. Number of operational modes available. An operational mode is a means of implementing a system function. For example, mine detection in the SMS system may be accomplished wholly (automatically) by the computer; by the operator with the aid of the computer; or entirely manually. A system with alternative operational modes obviously has greater flexibility and capability to respond than a system with only one mode. Number of operational modes is obviously correlated with system complexity but may also improve the quality of system output.

3. Number of subsystems. System complexity is related to but not completely determined by this parameter. Number of subsystems probably increases number of communications channels and resultant communications.

4. Subsystem organization. Some systems are organized hierarchically (in series), i.e., outputs from lower levels are transmitted to higher levels. Other systems are organized laterally and subsystems function in parallel, transmitting their outputs more or less concurrently to a higher level. Many systems combine both types of organization. A series organization is one in which each subsystem is dependent on the adequacy of immediately previous subsystem functioning; consequently in such systems the potential for breakdown is increased.

5. Number and organization of positions/personnel. Systems obviously vary in the number of positions that must be filled. This parameter is also related to system complexity but it is unlikely that systems with more personnel automatically become more susceptible to degradation of performance quality.

Personnel organization defined in terms of the roles played by system personnel may also be important. Associated with the hierarchical structure of subsystems, certain positions may be considered as "key" (especially important) positions at which information is received, filtered, interpreted, used as a basis for decision-making, etc. Key positions probably have more impact on output than do others. Systems will vary in terms of the number of "key man" positions.

6. Number, type and locus of transforms necessary for a system output. If each transform represents a critical transformation within or between individuals, teams or subsystems, one might hypothesize that the more of these, the more probable it is that errors, critical delays, lack of output quality, etc. will occur. Whether this happens may be dependent on the type of transform (whether it is decisional, interpretative or communicative) and where in the system hierarchy it occurs. From a measurement standpoint the transform indicates a point at which personnel performance data should be taken.

7. Number and organization of communications channels. Communications channels are particularly important to system functioning because, outside of matter/energy inputs/outputs, the only inputs/outputs transmitted within and beyond the system are in the form of information which must be communicated. Miller (1978) suggests that "there is always a constant systematic distortion between input and output of information in a channel" (hypothesis 3.3.3.2-2, p. 96) and that "a system never completely compensates for the distortion in information flow in its channels" (hypothesis 3.3.3.2-4, p. 96). His following hypotheses (selected from others) may be particularly relevant to manned systems:

"The probability of breakdown of adjustment processes among subsystems of a system decreases as the number of parallel information channels serving it increases" (hypothesis 3.3.3.2-10, p. 96).

"The probability of error in or overload of an information channel is a monotonic increasing function of the number of components in it" (hypothesis 3.3.3.2-11, p. 97).

"The less decoding and encoding a channel requires, the more it is used" (hypothesis 3.3.3.2-16, p. 97).

"The information input with the greatest intensity or greatest signal-to-noise ratio is given priority processing" (hypothesis 3.3.3.2-19, p. 97).

"A system gives priority processing to information which will relieve a strain (i.e., which it "needs"), neglecting neutral information. It positively

rejects information which will increase a strain." (Hypothesis 3.3.3.2-20, p. 97).

"In periods of stress and/or change in a system, the amount of information processing relevant to both task performance and adjustments among subsystems increases" (Hypothesis 3.3.3.2-21, p. 97).

"As the amount of information in an input decreases (i.e., as it becomes more ambiguous), the input will more and more tend to be interpreted (or decoded) as required to reduce strains within the system" (Hypothesis 3.3.4.2-9, p.99).

Communications channels are an obvious locus for performance measurement.

8. Output requirements. Does the system have to produce a specified number of outputs? As many outputs as possible? Are these outputs quantitatively specified? How much output variation is permitted (e.g., the range of acceptable variation)? To the extent that output variation is specified and tightly controlled, precise system feedback and evaluation are possible. One can also measure output quality in terms of the requirements imposed on it.

9. Characteristic inputs. Is it possible that systems can be differentiated by the nature of the inputs that activate their mission? What type of inputs activate the system (e.g., matter/energy or information)? Are these a single type or does the system react to different types? How frequent are these inputs; are they persistent during system operation? Do system operations change the nature of these inputs? Again the investigator may wish to examine output quality in terms of the kinds of inputs the system receives (see also remarks about indeterminacy).

10. System reactivity. Is the system reactive (i.e., initiates its mission only when stimulated by external inputs) or does it seek out the object of its functions based on its own internal programming? This parameter is obviously related to the types of functions the system performs and may be too gross to be related directly to system output. However, the system's activity cycle

may differentiate system types. Is that activity continuous or episodic? Are there periods when the system appears dormant because it is engaged solely in self-maintenance activities (e.g., housekeeping, maintenance, training)? Does it have many functions like emergency procedures that are only rarely implemented? The sequence of system activities over time and in pursuance of particular functions may permit us to categorize systems, albeit very grossly.

11. Degree of mechanization. Obviously systems vary in the extent to which they depend on mechanization. Logically it would seem that purely manual systems should be slower and more subject to error or quality degradation than more mechanized systems; but controlled tests of this hypothesis have not to the author's knowledge been performed. In any event, our natural pessimism makes us suspect that the relationship between mechanization and system performance is probably indirect, mediated by other variables (which have not yet been identified).

12. Feedback to system personnel. Systems probably also vary in the extent to which they provide feedback about mission events/success to system personnel. Based upon what is known about feedback effects on individual performance (see Meister, 1976) one could hypothesize that the more feedback the more effective system outputs should be, since feedback permits the system to monitor and to control system quality. Again, however, we suspect that all sorts of mediating variables affect this correlation.

13. System indeterminacy, as defined by Katz (1974) and Meister (1975). Indeterminacy in systems is composed of three variables: (1) the nature of stimulus inputs; (2) the amount of flexibility permitted to personnel in implementing system operations; and (3) the degree of personnel response programming. Thus a system in which inputs are ambiguous or capable of multiple interpretations based on molecular stimulus differences; in which alternative procedures are available (determined by the meaning assigned to these inputs); and in which personnel responses vary as a function of the preceding two variables, can be considered highly indeterminate. A system in which inputs are invariant (or almost so); in which only one procedure is necessary or

permitted; and which is capable of only one or two personnel responses is highly determinate.

For example, a missile launch control operation is highly determinate because inputs (console indicators) have a relatively unambiguous meaning (satisfactory, out of tolerance, emergency); the operating procedure calls for simple activation of a series of switches in a prescribed order; failure to perform a required step in the prescribed sequence prevents the launch from proceeding. An example of a highly indeterminate system is a divisional headquarters during an attack where inputs are fragmentary and often ambiguous, the order of battle must be progressively built up and the successful response (which will counter the enemy) is a probability only.

Of all the parameters listed above, the author's favorite is indeterminacy, in part because it is comprehensive (involving input, output and mediating processes) and may subsume a number of the other parameters. Moreover, indeterminacy can be operationally defined more readily than some of the others and can be scaled on a continuum from almost completely determinate to almost completely indeterminate. In consequence one can derive testable hypotheses from it. As an exercise to see what hypotheses system research can deal with, it will be useful to consider the hypotheses that one can derive with indeterminacy as the independent variable. No claim is made that the following is either comprehensive or indeed other than exemplary.

1. With increasing indeterminacy, the number of operational modes available to the system increases. Rationale: with a variety of stimulus inputs it is likely that different ways of responding to these inputs will be necessary.

2. The amount of communications processed by the system increases with increased indeterminacy. Rationale: highly variable inputs require more processing which is implemented primarily through communications.

3. The amount of feedback provided to system personnel has a direct relationship to the amount of indeterminacy in the system. Rationale: more system

adjustments are required in indeterminate systems, hence more monitoring of progress.

4. The amount of decision-making (and correspondingly the number of transforms) increases with increases in system indeterminacy. Rationale: the number of alternatives (ambiguous inputs, alternative responses) increases with indeterminacy; selection among alternatives requires decision-making.

5. The more indeterminate the system, the greater skill and more training system personnel must possess. Rationale: increased decision-making is required in indeterminate systems (because of the number of alternative responses available), hence more skilled personnel are needed to make those decisions.

6. The probability of information overload increases with amount of system indeterminacy. Rationale: in an indeterminate system more information must be solicited and accepted to interpret ambiguous and hence unstable input conditions.

7. The greater the degree of indeterminacy, the more impact personnel performance will have on system output. Rationale: with ambiguous inputs, partial information, etc. system outputs cannot be inflexibly programmed but must be channeled through interpretive processes controlled by personnel. Under these circumstances the potential for human error increases and consequently the probability of output degradation. Of course, the relationship between indeterminacy and personnel impact is not really that simple, since skilled personnel can often compensate for error and degraded conditions.

The above are almost certainly not the total number of relationships that can be derived between indeterminacy and other system variables. They are however all the author could think of; lacking empirical data about variables inherent in system functioning, one cannot allow one's imagination to run riot. Certain relationships seem inherently improbable, e.g., indeterminacy and (a) number of positions/personnel; (b) type of function performed (although it would be interesting to hypothesize that surveillance and decision-making functions are more closely associated with indeterminacy); (c) number

of subsystem levels (although indeterminacy may build up with increasing subsystem echelons); (d) degree of mechanization.

All of this is gut feel, however. We simply do not have enough information about actual systems to do more than make very tentative hypotheses. However, the listing of potential system parameters does provide one way of securing information about systems. One can attempt to classify operational systems in terms of the above listing and look for relationships with system outputs. New parameters and relationships may thereupon suggest themselves.

The significance for PSM of the indeterminacy construct is its impact upon the measurement strategy adopted. For example, because there is little stimulus-response variability in these systems, highly determinate systems should have more precise criteria and standards of performance, which make it easier to evaluate personnel performance. Such systems can be more readily evaluated by use of so-called objective measures; in indeterminate systems, since the meaning of inputs and responses depend in part on personnel interpretation, it becomes necessary to use more subjective instruments (e.g., interviews), to determine what that interpretation is.

Again, because of their relative lack of variability, it is easier to simulate determinate systems and hence to measure their performance in simulators.

The parameters and hypotheses advanced so tentatively can be tested in two ways, operationally and experimentally:

1. Operationally one would examine a variety of "real world" systems in their environment and attempt to describe their functioning in terms of the parameters and hypotheses presented previously. Concretely this means collecting a great deal of descriptive data about such variables as inputs, functions performed, procedures for dealing with inputs and outputs, communication channels, organizational relationships among subsystems, changes in system responses over mission time, etc. Observation of on-going activities, interviews with key personnel, written questionnaires and rating scales and

examination of available documentation would be necessary. This procedure would not require the collection of human performance data except in a very molar manner. Analysis of the resultant data would be essentially correlational, since manipulation of variables would hardly be possible. Hypotheses would direct the investigator to the specific variable relationships to be tested, with special attention to system outputs.

2. Experimentally the researcher would develop an "analogue system" containing the parameters he wished to test. For example, he might model an air surveillance system with inputs presented in the form of visual stimuli whose characteristics (in terms of the responses required of subjects) ranged from unequivocal to extremely ambiguous. The researcher would specify the responses to be made to these stimuli (e.g., categories such as missile, aircraft, natural phenomena) with procedures for resolving ambiguities. If he wished to determine the effect of indeterminacy, he would construct problems which varied in amount of indeterminacy and, say, number of personnel or feedback as independent variables and speed and accuracy of terminal output decision-making as dependent variables. Each subject, exposed to the set of problem situations, would serve as his own control.

Note that what has been described is classic CE which, as was pointed out before, is entirely appropriate for system research as long as the requirements of system research (see Chapter Two) are satisfied.

The two approaches are complementary, not opposing. Operational studies should be initiated before experimental ones because, being closer to the real world, they are more likely to produce leads that can be turned into hypotheses for more precise experimental test. Such operational studies should moreover continue even as experimental ones are proceeding. Any conclusions reached experimentally should be brought back into the operational environment for verification with actual systems. If there is a system research paradigm, therefore, it is a progression from the operational environment to the laboratory study, back to the operational environment.

CHAPTER SIX

CRITERIA FOR SELECTING AND EVALUATING PSM RESEARCH

In this chapter we examine the question: Are the criteria ordinarily applied to the selection and evaluation of scientific research--validity and reliability--adequate for PSM research, and, if not, what other criteria apply? From the way the question was phrased, it is obvious that we have answered it in the negative, and so we will have to explain why.

Any discussion of research presupposes a value judgment: research possessing certain characteristics has greater worth (however worth is defined, and this is an individual matter) than research possessing other characteristics. Criteria have always been applied to research, but generally those accepted by the scientific community have described only validity and reliability. Other criteria exist, but have rarely been considered. It is the author's point of view that the restriction to validity and reliability is associated with much behavioral research having only limited usefulness to the Human Factors discipline.

Research criteria should be applied at two stages in the research process:

1. Before a study is initiated, the researcher uses these criteria to decide whether or not he should proceed with that study.
2. After a study is completed, the researcher (and others) use these criteria to evaluate the worth of the study and its results.

Our particular interest lies in the former application, because criteria are far more valuable when used to select the studies to be performed than they can ever be in evaluating those studies after they have been completed. Once completed, a bad study clutters up the journals and misleads readers; it is, therefore, far more important to prevent the bad study from being started. For this reason it is important for the reader in examining the criteria discussed below to consider how effectively a particular criterion can be used before a study is performed.

In addition to the traditional validity and reliability, we will consider: relevance; applicability; generalizability; and utility.



1. Validity. Everyone presumably knows what validity is, although actually most do not and the subtleties involved in the concept would make a rabbi's eyes bulge. There are different validity concepts, some expressed in purely statistical terms, others in philosophical ones. Our definition of validity is phrased in the form of a question: Do the research results truly represent what one has measured or hoped to measure? The key word is "truly," because at the heart of the validity criterion is the concept of some absolute standard of truth, so that if X is an event, phenomenon or behavioral state, X' (the results of the measurement) equals or approximates that X.

A discussion of the complexities of the validity criterion is outside the scope of this monograph, but it is necessary to point out that in an absolute sense, validity can never be established because it presupposes a standard of comparison (between measurement results and the event, object or phenomenon being measured) that is independent of measurement operations. Unfortunately, the researcher's knowledge of an object, event or phenomenon can never be completely independent of the processes by which he measures that object, event or phenomenon. Because inaccuracy is inherent in all measurement processes, validity cannot be absolutely ensured.

The researcher can, however, attempt to gain confidence in his conclusions in various ways: (a) by predicting performance in another situation based on the results of the earlier measurement (predictive validity); (b) by repeating the measurement with a different set of subjects, possibly under different conditions, and noting whether the same results are attained (convergent validity). Other types of validity, like construct validity, are essentially mythical.

If one looks at measurements of relatively molecular behavioral phenomena (e.g., reaction time to discrete stimuli), predictive and convergent validity coefficients are fairly high. As study variables become more molar, however, it becomes progressively more difficult either to predict future performance or to repeat the results of earlier studies.

PSM makes more extensive use of predictive validity because many of its measurements are performed to solve problems or decide upon a course of action. If the problem is solved or the action is effective, there is at least a suggestion of predictive validity.

The use of validity as a means of deciding which study rather than another should be performed (because the data from that study will be more valid) is in practice not feasible; because it requires the researcher to anticipate the relationship between a future event--data from a study not yet performed--and an actual event or phenomenon, the truth about which cannot be ascertained. This is too difficult for anyone. In fact no researcher cudgels his brain in that fashion. In practice validity as a research selection criterion is pleasantly ignored, as it is also (to a somewhat lesser degree) in research evaluation. The use of validity as an evaluation criterion (to ascertain in which studies one should rest confidence) is shot full of the researcher's biases, because, lacking any objective basis, it permits him to ignore study conclusions that conflict with those biases.

If validity is of dubious utility in CE, one would not expect it to be of much greater use in PSM. And in fact the PSM researcher assumes validity as does everyone else. His assumption has a more solid foundation, however, because the measurement sources he deals with are closer to operational reality.

2. Reliability is the consistency of the researcher's results, i.e., repeated measurements on the same object, event or phenomenon, performed in the same manner each time, will produce essentially similar results. For measurement situations which can be repeated more or less identically with the same subjects (e.g., a standard test, a highly controlled experimental situation) the term has meaning. Because, as we suggested previously, (a) the mission scenario is the control element in PSM (in operational studies, anyway); (b) individual scenarios can vary substantially (although still within their broad, common structure); and (c) opportunities for the same subjects to repeat task performances in the same scenario are often lacking, it may be somewhat more difficult to establish the reliability of PSM than of CE data.

Reliability, like validity, is almost never used as the basis for selecting one measurement situation over another, because it is difficult to anticipate data consistency except in very general terms. However, reliability comes into its own as an evaluation criterion. PSM finds the reliability concept to be useful but may have some difficulty in utilizing it because of the reduced control under which some PSM measurements are made.

3. Relevance indicates whether measurement results relate to the questions/purposes for which the study was initiated.

The importance of this criterion should be readily apparent. Research always begins with a question/purpose which, at least by implication, imposes a requirement upon the research: to answer that question or accomplish that purpose. In a sense research purpose is equivalent to system mission.

Often that question/purpose is very specific: for example, does variable X have a significant effect on a particular performance? The more specific the question/purpose, the more easily one can determine whether the results secured are relevant to that question or purpose. Similarly, the more specific the question, the more likely it is that the study results will be relevant, because the precision of the question makes the selection of an appropriate methodology more probable.

There is, however, a hierarchy of questions/purposes. Behind the immediate, specific question is always a more molar one, which is really what the study is designed to achieve. Because of this, the answer to the more specific question implements the higher order one. Example: Immediate study purpose: to determine the finger pressure required to activate two and three pole switches. Higher order study purpose: to develop data permitting the design engineer to select the most efficient controls for operator use. Obviously the lower order purpose/question is "nested" in the higher order one.

In order to determine the relevance of any study it is necessary to ask what its higher order purpose is and how closely the specific study purpose corresponds to the higher order one. Relevance is always in terms of that higher order purpose. It is not permissible, however, to specify as a higher order purpose such generalities as "science" or "basic knowledge," because almost anything can be subsumed under such generalities.

Although the specific study question/purpose is generally articulated very carefully, very little consideration is usually given by the researcher to its higher order relation. Paradoxically, it is entirely possible to perform a study that satisfies the more specific purpose while failing to satisfy its higher order one. That is because there may be so great a distance between the

two orders of purpose that the study cannot bridge the chasm. Suppose, for example, one did a study on expectancy theory although the ultimate purpose of that study was to contribute to human engineering or system design principles. It might be difficult to see the relatedness of the two purposes. This is the case with much behavioral research in which it is difficult to see what ultimate purpose that research serves. One need only examine journals like the Journal of Experimental Psychology to see that little consideration has been given the relevancy criterion.

The more general the higher order study purpose, the more judgment enters into determining the relevance of measurement results to that purpose. Hence there is room for honest disagreement about whether a particular study makes a contribution to the higher order purpose implicit behind the specific question being answered. It would be desirable to make researchers specify the higher order purpose of their work and to show the relationship between their study results and that purpose. Whether this would in fact reduce the number of irrelevant studies cannot be predicted.

One might expect that researches performed by Human Factors specialists specifically for Human Factors purposes would be more relevant to those purposes than research of a general nature which is adapted to Human Factors needs. Often one finds, however, that the two categories of research are equally irrelevant, which suggests that much research supposedly performed for specific Human Factors purposes is actually performed under disguise, as it were.

The relevancy criterion can and should be used to create the most appropriate study or the most appropriate study methodology (for example, making a decision between collecting data in the laboratory or the operational environment). Relevance, unlike validity and reliability, is feasible as a decision criterion because the researcher need only relate his study methodology and the kind of data he intends to collect to the higher order study purpose. The system oriented researcher can do this more readily than the CE researcher because his orientation is to the operational system (which is, to say, to reality).

4. Applicability indicates the degree to which measurement results may be transformed into actions solving a problem or enabling a prediction to be

made. The essence of this criterion is that there are action consequences of the research.

It is generally accepted that the purpose of research (i.e., science) is to gain more adequate understanding of the world (i.e., system operations). But understanding unsupported by actions based on that understanding is specious, being too often manifested merely by ignorant theorizing. Volumes of behavioral research are filled with windy words with which one can do nothing. Such understanding creates pleasure in those who think they understand, but nothing more. Only when one can apply these words can one speak truly of understanding.

Applicability, again like relevance, is a relative criterion. There is no absolute standard of applicability to which every study must conform. Its major usefulness is in deciding between alternative study projects or methods of measurement.

Manifestly, PSM research is more applicable than traditional CE, because PSM is more action-oriented than CE. Applicability, like relevance, can be readily used as a means of deciding what and how a study is to be performed.

5. Generalizability indicates the degree to which measurement results can describe objects, events or phenomena similar to but not identical to those on which the measurements were made. Some generalization is necessary in all research because any study is based on a sample of the population which differs to some extent from the parent population. The more generalizable a set of measurement data (or the conclusions from these), the more valuable the research is.

Generalizability is less important than relevance and applicability: It can, however, be added to the latter criteria when a decision must be made among alternative study methods. In actual practice it is rarely used for this purpose. It is occasionally used as an evaluation criterion.

6. Utility can be defined in terms of three dimensions:

a. The criticality of the problem to be solved or the question to be answered.

b. The amenability of the problem or the question to measurement processes.

c. The possibility of applying the measurement results in the real world (not whether they will).

Like the previous criteria, the utility criterion should be applied only in deciding among alternative studies. There is no way of applying an absolute standard of utility; what is a critical subject for one researcher may be trivial to another. Once a subject area has been selected, however, the researcher can choose rationally among alternatives within that subject.

A study has high utility if the problem/question attacked is relatively important (in terms of impact upon the system = the real world); if the problem/question can be measured reasonably effectively; if the measurement results can actually be utilized in performing some action or developing a consequence related to the problem/question.

A study would not appear to be worth performing if (a) the problem is unimportant; for example, there would seem--to this author at least--not much point to studying toilet graffiti even if one were a sociologist interested in studying scatology; (b) the problem/question does not lend itself to measurement; for example, when the author worked for the Army a research topic was several times proposed (by operational personnel) that we considered not amenable to measurement: to study methods of making combat less stressful for troops; (c) the study results cannot be applied meaningfully to the real world situation from which the problem/question originally arose; for example, any research results dealing with making combat less stressful would be unlikely to be effective when real bullets are used.

It is apparent that the utility criterion can be more easily applied in PSM research than in traditional "fundamental" research.

This list of criteria is technical only. There are factors such as cost, the effort involved in performing the study, the acceptability of the research topic to one's peers and superiors and practical problems of implementation, that must also be considered. However, these are outside our purview.

Several points should be noted. The criteria described are not completely independent of each other. Utility and applicability are interwoven; so are relevance and utility. On the other hand, validity and reliability are rather independent of the others, although not of each other: one can perform a highly valid and reliable study which is neither relevant, applicable, generalizable or useful. (Much behavioral research is of this nature.) The other four criteria have real world referents whereas validity and reliability have no reference other than their own measurement processes. (The phrase "referents" refers to elements that must be taken into consideration in applying the criterion.) This may account for the popularity of these two criteria in purely academic circles. Another reason for their popularity is that validity and reliability can be expressed quantitatively, as coefficients of correlation. The others are judgmental only. On the other hand, validity and reliability can be applied only post facto, whereas the others can also be used to select the potentially more effective study among alternatives.

Ideally the investigator wishes to have both valid and reliable measurements and is content when he feels he has accomplished these. But is this sufficient? Valid and reliable measurements may well be irrelevant to the questions with which the study began. Suppose, as a purely hypothetical example, a study seeks to assess the physical prowess of 50 year old men, but the measure taken counts only the number of "pushups" they can do. The measure is a valid measure of pushups and is highly reliable; however, the data describe only one aspect of strength, and, as a consequence, they are only partially relevant. The difficulty lies in failing to correlate the measurement questions asked with an instrument appropriate to those questions. The more complex the questions, the more likely such a failure of coordination is to occur. The problem here is analytical rather than mensurational: it occurs because the investigator fails to describe his study goals precisely enough in advance of measurement, and, as a consequence, he selects an inappropriate instrument.

Beyond validity, reliability, and relevance are the results capable of being transformed into some consequence or action? Literally, what can one do with the effects of measurement? It is undeniable that knowledge is its own goal; but how much better is that knowledge when it can be used for something? Besides, there are degrees of knowledge value; certainly not every study in the vast outpouring of information in the behavioral sciences is equal to every

other study. And if that is the case, does it not argue that those studies that have action consequences are better than those that do not--always assuming, of course, equal validity, reliability and relevance.

Even if one has valid, reliable, relevant and applicable results, these may or may not be generalizable to objects or events other than the ones measured. Generalizability is not as important as the preceding factors; it limits measurement utility but it does not absolutely destroy it as would lack of validity, reliability or relevance. Generalizability depends on the extent to which the object or event measured resembles other objects or events in terms of the dimensions being measured. This is a function of the dimensional range of the subject sample. The more specific the measurement, the less it can be generalized. If one measures the strength of 12 year old girls, the results can be generalized only with difficulty to 40 year old women. The characteristics of the object or event being measured determine its generalizability; if one wishes to expand generalizability, the range of those characteristics must be expanded.

If one had an interval scale to measure it, the utility of data for the satisfaction of the measurement purpose could vary from zero to theoretical infinity. Each of the variables above adds to (or subtracts from) utility. Lack of generalizability (to a large enough population) reduces utility slightly; lack of applicability reduces it much further; invalidity, unreliability and irrelevance totally destroy it.

The point of this chapter is that the investigator cannot be satisfied merely if his measurement is valid and reliable, the usual criteria for evaluating a study. Moreover, he should utilize these criteria not merely for the evaluation of a study after it has been completed (and then usually only for studies performed by others), but before he begins his own work. Although the system concept does not strictly require the application of relevance, applicability, generalizability and utility criteria, the system orientation is more conducive to their use than more traditional CE.

REFERENCES

- Askren, W. B. and Lintz, L. M. Human resources data in system design trade studies. Human Factors, 1975, 1, 4-12.
- Askren, W. B. and Newton, R. R. Review and analysis of personnel subsystem test and evaluation literature (AFHRL-TR-68-7). Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, January 1969.
- Bavelas, A. Communication patterns in task-oriented groups. Journal of the Acoustic Society, 1950, 22, 725-730.
- Bourne, L. E. Effects of delay of information feedback and task complexity on the identification of concepts. Journal of Experimental Psychology, 1957, 54, 201-207.
- Bryan, G. L. et al. The AUTOMASTS: An automatically-recording test of electronics troubleshooting (Tech. Rep. 11). Electronics Personnel Research Group, University of Southern California, Los Angeles, California, August 1954.
- Buckner, D. N. and McGrath, J. J. (Eds.) Vigilance: A symposium. New York: McGraw-Hill, 1963.
- Chapanis, A. The relevance of laboratory studies in practical situations. Ergonomics, 1967, 10, 557-577.
- Damrin, D. E. and Saupe, J. L. Proficiency of Q-24 radar mechanics: IV. An analysis of checking responses in troubleshooting on tab test problems. Air Force Personnel and Training Research Center, Lackland AFB, Texas, November 1954.
- Grimsley, D. L. Acquisition, retention and retraining: Effects of high and low fidelity in training devices (Tech. Rep. 69-1). Human Resources Research Office, 1969.
- Grings, W. W. et al. Shipboard observation of electronics personnel: Detailed descriptions of observational techniques (Tech. Rep. 2). Electronics Personnel Research Group, University of Southern California, Los Angeles, California, January 1953.
- Hovland, C. I. Human learning and retention. In Stevens, S. S. (Ed.), Handbook of Experimental Psychology, New York: Wiley, 1951.
- Hughes Aircraft Company. System design characteristics and user skills: An engineering guide for Navy electronics systems (Final Report, Contract N00123-77-C-0782). Hughes Aircraft Company, Fullerton, California, 1978.
- Katz, F. E. Indeterminacy in the structure of systems. Behavioral Science, 1974, 19, 394-403.
- Kidd, J. S. A summary of research methods, operator characteristics, and system design specifications based on the study of a simulated radar air traffic control system (Rep. 59-236). Wright Air Development Center, Wright-Patterson AFB, Ohio, July 1959.

- Mackie, R. R. et al. New criteria for selection and evaluation of sonar technicians: Phase II. Trial administration of experimental predictor tests (Tech. Note 78-13). Navy Personnel Research and Development Center, San Diego, California, May 1978.
- Meister, D. Methods of predicting human reliability in man-machine systems. Human Factors, 1964, 6, 621-646.
- Meister, D. The indeterminate man-machine system. Proceedings, Human Factors Society Annual Meeting, October 1975.
- Meister, D. Behavioral foundations of system development. New York: Wiley, 1976.
- Meister, D. Implications of the system concept for Human Factors research methodology. Proceedings, Human Factors Society Annual Meeting, October 1977.
- Meister, D. Subjective data in human reliability estimates. Proceedings, Annual Reliability and Maintainability Symposium, Los Angeles, California, January 17-19, 1978, pp 380-384.
- Meister, D. and Rabideau, G. F. Human factors evaluation in system development. New York: Wiley, 1965.
- Meister, D. et al. The impact of manpower requirements and personnel resources data on system design (AMRL-TR-68-44). Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, September 1968.
- Meister, D. et al. The effect of amount and timing of human resources data on subsystem design (AFHRL-TR-69-22). Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, October 1969. (a)
- Meister, D. et al. The design engineer's concept of the relationship between system design characteristics and technician skill level (AFHRL-TR-69-23). Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, 1969. (b)
- Meister, D. et al. Relationship between system design, technician training and maintenance job performance on two autopilot systems (AFHRL-TR-70-20). Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, September 1971.
- Miller, J. G. Living Systems. New York: McGraw-Hill, 1978.
- NAVTRAEQUIPCEN. Analysis of the transfer of training substitution, and fidelity of simulation of training equipment (TAEG Rep. 2). Naval Training Equipment Center, Orlando, Florida, 1972.
- Parsons, H. M. Man-Machine system experiments. Baltimore: Johns Hopkins Press, 1972.
- Peters, G. A. and Hall, F. S. Missile system safety: An evaluation of system test data (ROM 3181-1001). Rocketdyne, Canoga Park, California, 1 March 1963.
- Prophet, W. W. and Boyd, H. A. Device-task fidelity and transfer of training: Aircraft cockpit procedures (Tech. Rep. 70-10). Human Resources Research Office, Ft. Rucker, Alabama, 1970.

Steinemann, J. H. Evaluation of graduates of the electronics technician Phase A-1 training program (Rep. SSR 68-20). Naval Personnel Research Activity, San Diego, California, March 1968.

Streuning, E. L. and Guttentag, M. Handbook of evaluation research. Beverly Hills: Sage Publications, 1975.

Tanner, W. P. and Swets, J. A. A new theory of visual detection (Tech. Rep 18). Electronics Defense Group, University of Michigan, Ann Arbor, Michigan, 1953.

Van Cott, H. P. and Kinkade, R. G. (Eds.) Human engineering guide to equipment design (revised edition). Washington D. C.: Government Printing Office, 1972.

Williams, H. L. and Malone, J. S. Evaluation of the 3-M system as implemented by the naval surface forces in the San Diego area (Rep. 78-12). Navy Personnel Research and Development Center, San Diego, California, June 1978.

APPENDIX I

CRITICISMS OF THE SYSTEM APPROACH TO MEASUREMENT

APPENDIX I

CRITICISMS OF THE SYSTEM APPROACH TO MEASUREMENT

The author has heard a number of colleagues raise the following objections to the approach described in this monograph. These objections center around the following points.

1. The criterion of utility is not relevant to science.
2. There is nothing new in the system approach and Human Factors specialists already accept its premises.
3. The over-concentration on the manned system breaks completely with Psychology.
4. There are inadequacies in present Human Factors methods, but things are not that bad and time and more research will solve these problems.
5. It would be highly desirable to do research as the system approach suggests, but it is too difficult.

In the remainder of this Appendix, we will review these criticisms and see whether they are well founded.

1. The criterion of utility is not relevant to science.

The utility criterion is implicit in the system approach because the focus of that approach is on the operational system which emphasizes immediate problems. Because of this the reader may believe that what we are talking about is merely application of more general principles (already known, of course). Later we shall make a case for the fact that Human Factors is not merely the application of experimental psychology but here we need concern ourselves only with the notion of utility.

The rallying cry of basic research is knowledge for its own sake. Presumably we cannot know the ultimate value of any study, no matter how apparently trivial.

A hundred studies of little importance may have to be performed before one of significant value is found. This is the philosophy that has driven most behavioral research (at least that part of it which is academically oriented and which one finds in journals like Human Factors and Ergonomics).

Research is manifestly not performed at random. Of the thousands of studies that could be performed on any subject, only a few are actually implemented. Obviously some judgmental criteria (see Chapter Six) were used to select those that were initiated. Implicit in those choices was some sort of utility criterion, e.g., the researcher's interest in a subject; the need to satisfy an academic requirement like the doctoral degree; the feasibility of performing the study; estimates as to relative payoff. None of these criteria is purely technical; most are simply personal.

It is myth, therefore, that the researcher in performing basic research has complete freedom to do what he chooses to do. Even if he had almost limitless freedom, he would still be constrained by his own concepts of relevance.

It is apparent, therefore, that some utility criterion--however it is phrased--is implicit in the researcher's choice of a study to perform. The criteria applied by the basic researcher are not qualitatively different from the ones suggested in this monograph; they are merely less immediate. Basic research is that research which has a payoff only in a more or less remote future, whereas applied research has a goal which can be more concretely imagined. If this is once granted, why not apply the criteria implicit in the system concept?

2. There is nothing new in the system approach and Human Factors specialists already accept its premises.

If the system approach is really accepted by most Human Factors researchers, it is difficult to see why the research they perform as reported in the journal literature lacks that orientation. The implication of the system concept as far as measurement is concerned is that studies performed on individualistic variables in a non-operational environment must be verified by studies performed on those variables in operational systems. Consequently a major goal of Human Factors research should be to determine the applicability of psychological research to the system characteristics of Human Factors.

One sees, however, little if anything in the literature involving systems as systems and the system approach as discussed in this document. One cannot accept the system approach and simultaneously ignore its implications. If those implications are accepted, it is necessary, for example, to question the behavioral research conclusions so far developed because none has been verified with reference to operational systems. It is necessary to ask why we have so little data about manned systems and the role of the human in those systems.

The man-machine system as it has been described for example in Chapter One of Van Cott and Kinkade (1972) is not the system as we have described it here. The system cannot be limited (as it usually is) to the immediate man-console interaction. Even if it were so limited, the manner in which system research on that interaction should be performed is far different from the way in which man-machine research is currently performed.

It is, of course, possible that the system approach is accepted by Human Factors specialists on a purely abstract theoretical basis, but that the concrete implications of that approach are ignored because specialists are more comfortable with the research strategies they learned in school. Those strategies, it need hardly be emphasized, reflect an orientation toward the individual, not the system.

3. The over-concentration on the manned system breaks completely with Psychology.

This criticism is somewhat justified. Since the subject matter of Human Factors is the manned system, the principles, techniques and conclusions derived from psychological research which is oriented toward the individual cannot be applied uncritically to the system.

Manifestly the manned system includes the human and, therefore, there is a continuity between Human Factors and Psychology. That is why we emphasize that the conclusions derived from psychological research may apply to the system; but since in the manned system we are dealing with a level of discourse qualitatively different from that of the individual, it is necessary to test these psychological conclusions against the operational system. It is possible that:

a. Some of these conclusions (perhaps the majority) will apply to personnel performance in the manned system wholly and without modification.

b. Some of these conclusions will apply but will have to be modified in ways that we cannot presently envisage.

c. Some of these conclusions will have minimal impact in the system context and, therefore, can be discarded (but only as they relate to the manned system).

d. Entirely new principles not presently conceived may have to be developed.

Until we test our present psychologically-derived conclusions against the reality of the manned system, we cannot say which conclusions are valid for the system and which are not.

We have implied in the foregoing that the manned system is something entirely different from the individual and, therefore, that the individualistic orientation of psychology does not apply at the level of the system operator. This is not to suggest that the individualistic orientation is invalid but merely that it does not apply to the system situation. For example, what is the significance of principles of electrical brain activity to the performance of a technician attempting to maintain a malfunctioning equipment? How does eyelid conditioning relate to the student pilot learning to land on a carrier? Manifestly, electrical brain activity is necessary for the technician to function; but it does not serve to explain his troubleshooting behavior. We know also that we can condition the eyelid responses of the student pilot; but this is irrelevant to his learning a complex psychomotor landing pattern.

By our reasoning Human Factors specialists are not psychologists, although their initial training may have been in Psychology. Although one can recognize the Human Factors debt to Psychology, it is necessary also to recognize that the two disciplines diverge because their subject matters are different.

4. There are inadequacies in present Human Factors methods, but things are not that bad and time and more research will solve these problems.

One reason the researcher may have for looking at alternative methods of performing Human Factors research is that he is unhappy with progress in his field. If he is satisfied with the status quo, he is unlikely to look for other solutions.

How does one answer the objection that things are not so bad with Human Factors? After all, whether or not one is satisfied with its methodology depends on very subjective criteria.

The specialist's feeling of relative satisfaction should be determined by his answers to the following questions (which are merely exemplary):

a. Can one presently predict quantitatively how well an individual will perform in operating a control panel or console?

b. In evaluating the design of a man-machine interface, can one apply human engineering design principles quantitatively, e.g., the application of this design principle (arrangement of controls by function) will produce mean performance of X percent accuracy, whereas application of another principle (arrangement of controls by sequence of use) will produce X' percent accuracy)?

c. Does the designer or Human Factors specialist have at his disposal a series of tables which will allow him to predict personnel performance (in terms of accuracy, job completion or any other criterion) as a function of the following: type of man-machine interface; type of system/task; amount of prior training; skill level; motivation; etc.?

d. Does the training specialist have anywhere a set of principles that allow him quantitatively to determine how long a training program for a new design configuration should be and what that program should consist of?

e. Can we predict in the very early stages of system design what the effects on personnel performance will be of proposed changes in system design?

(We make the assumption, which not every Human Factors specialist may accept, that the goal of our discipline is to provide answers to questions such as these. These questions are oriented around the interaction of personnel performance with system design which we assume are the two immediate concerns of Human Factors. If the reader does not accept these assumptions, we have nothing to talk about.)

The answers to the above questions are largely negative. Can the Human Factors specialist then be satisfied with the progress of his discipline?

The commonly accepted reply is: No discipline fully achieves its goals; but that with time, money and research answers to those questions will be found. The Human Factors discipline is still comparatively quite young; surely another 25 or 50 years or . . . will give us what we want.

It is possible, however, that unless a more realistic approach to measurement is adopted, another 25 years of research will simply produce results not significantly different from the past 25 years.

5. It would be highly desirable to do research as the system approach suggests but it is too difficult.

Admittedly the approach described in this monograph is more difficult to pursue than our present measurement approach.

a. Working in the operational environment is more costly, more frustrating, requires more sensitive techniques and more creativity than working in the laboratory.

b. To verify in the operational environment results achieved in laboratory experimentation does involve additional burdens if we do not validate conclusions operationally.

c. Working with highly trained subjects does reduce our potential subject pool and requires that we spend much more time in training before we can experiment.

d. Simulating a system environment in the laboratory is more expensive in cost, time and effort than working on a purely individual level.

Having admitted all this, is it acceptable to use the difficulty of a new approach as an excuse for not getting on with that approach? Applying an incorrect strategy simply because it is easier to apply than a correct strategy is not an acceptable way of being a scientist.

APPENDIX II
SOME REPRESENTATIVE PSM STUDIES

APPENDIX II

SOME REPRESENTATIVE PSM STUDIES

In this appendix we present some "representative" PSM studies to illustrate the principles described previously. This is not a review of the literature, since only a few studies have been selected; for those interested in a broader survey we recommend Parsons (1972) and Meister (1976). The studies selected cover a broad time span, some of them having been performed in the early 50's, others so recently that their data are still being analyzed.

The studies were selected to illustrate the following categories of PSM measurement:

1. Exploratory tests
2. Resolution tests
3. Verification tests: Operational System Tests. (No examples of Operational Readiness tests could be found.)
4. Experimental field studies
5. Laboratory research
6. Normative data gathering

In reporting these studies we have endeavored to quote verbatim as much as possible from the original sources.

These studies illustrate the following points which exemplify PSM:

1. The impetus for Exploratory research is the absence of relevant answers in the available literature.
2. Much more use is made of subjective techniques in PSM than in laboratory studies.

3. With the exception of normative data gathering, all the tests described make use of CE principles, although the degree of control is less than in CE.

4. During system development Exploratory tests are often combined with Resolution and Verification tests.

5. In all of these tests the mission is critically important; as a result, operational fidelity is emphasized.

6. Special care is taken in selecting subjects with appropriate background and training. Subjects are often system personnel.

7. In studying operational systems either in the operational environment or in OST all aspects of the system (including training, procedures, technical data and system organization) are considered.

8. Since operational tasks are quite complex, measurement data can often be gathered and evaluated only by skilled operational personnel. Moreover, the more molar these tasks, the more difficult it is to use strictly objective, quantitative measurement techniques.

9. The length of many missions and the interaction of these missions with environmental conditions like weather often prolong operational testing far beyond the usual duration of laboratory studies.

1. Exploratory Tests

a. Personnel Subsystem Impact on C-5 Equipment and System Design, Parris, H. L. and Hall, T. J., Lockheed-Georgia Company, Marietta, Georgia (unpublished paper, no date).

(1) Crew Seat Comfort Study

Frequently, in the development of new aircraft, man-machine interface design problems arise where there is either no precedent or where extrapolation from the general case to the specific problem is tenuous or impossible. In these situations, special studies are required to obtain the data necessary for

formulation of definitive design recommendations. This paper summarizes two representative examples of real-world problems that required such special studies in the development of the C-5. Study methodology, principal results and conclusions pertinent to the effect on human performance as it relates to system performance are presented for: (1) the comfort evaluation of aircrew seats and (2) ground emergency egress provisions

A study was undertaken to systematically evaluate design characteristics of three alternative C-5 crew seats in terms of adequacy of human comfort as reflected by the physiological experiences of subjects in a simulated operational environment. A comparative approach was utilized. A C-141 seat served as the base line (Seat #1) from which the prototype of the C-5 crew seat was developed (Seat #2). After quantification of comfort on these two seats, results were used to modify the design of the C-5 crew seat (Seat #3) which, in turn, was then subjected to comfort evaluation.

Subjects sat in the respective seats for seven hours--four hours before lunch and three hours after lunch. The subjects were allowed to leave the seat to go to the rest room (adjacent to experimental area) and to obtain their meal at noontime

The method of evaluation consisted primarily of subjective tests administered by means of questionnaires presented to subjects

The basic criterion was a comfort rating scale administered at the end of the sitting period. Several other measures (hourly comfort evaluation, hourly prediction of the number of additional hours the subjects could sit in the seat, and hourly progression of discomfort in specific body regions) were used to reinforce the basic comfort rating and to identify any trends of comfort degradation during the sitting sessions.

The subjects were eight male employees of Lockheed-Georgia Company, selected to represent a range of anthropometric measurements

Subjects evaluated the seat each hour in terms of comfort/discomfort by checking one in a series of nine statements which ranged from a highly positive statement (+4) to a neutral statement (0) to a highly negative statement (-4). Results are given in Figure 1.*

*Original report Figure and Table numbers are used.

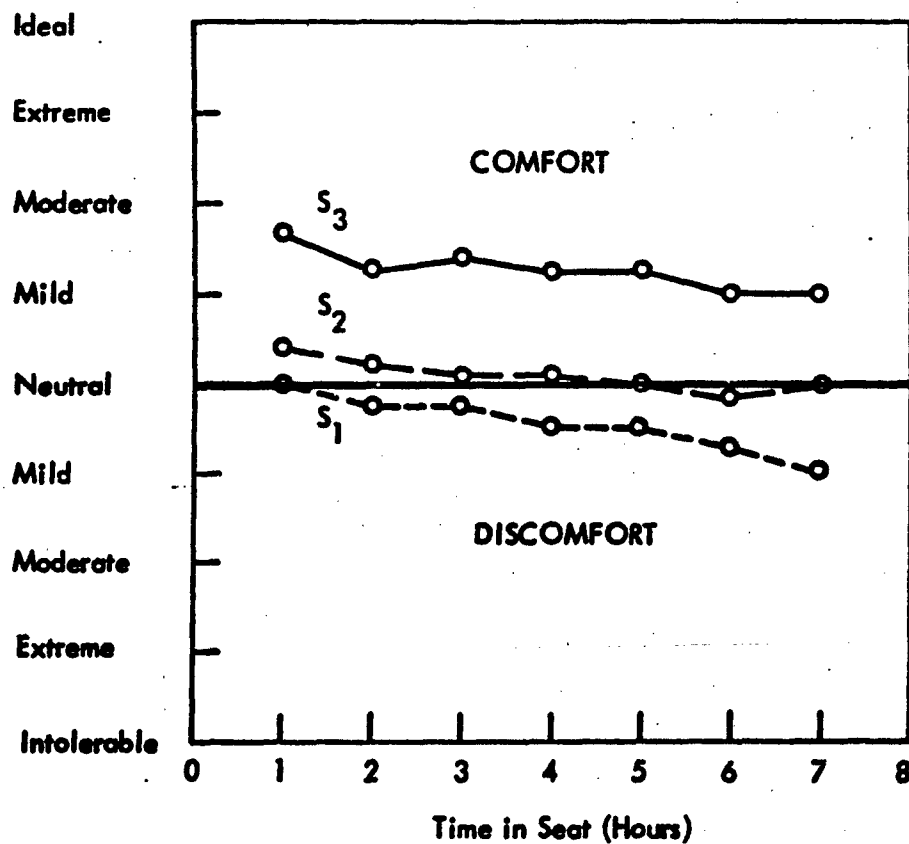


FIGURE 1. AVERAGE HOURLY EVALUATION OF COMFORT PROVIDED BY TEST SEATS

Seat #1 was typically assigned negative comfort values. Seat #2 was rated slightly comfortable during the initial hours of the test but gradually fell to a neutral comfort rating. Seat #3 was typically assigned positive comfort values, indicating that many of the uncomfortable features of Seat #1 and Seat #2 had been successfully eliminated.

The hourly predictions of additional number of hours that a given seat could be tolerated were averaged over subjects and depicted in Figure 2. The seats are clearly ranked for the first four hours with Seat #3 being the preferred seat for the first three hours with little difference between Seat #3 and Seat #2 after that point. After four hours, all seats cluster around seven hours of additional sitting time. Noteworthy is the fact that the group averages estimated that each seat could be tolerated for an additional six hours or more at the end of the seven hour testing session.

Comment: Additional data (not reported here) were gathered concerning the hourly progression of discomfort in specific body regions.

Note certain characteristics of this Exploratory test: (1) a prototype seat was fabricated and the information gained with this prototype was used to design the final seat; (2) although the Exploratory test does not require a comparative methodology, the latter is often used when an earlier design version of the system at issue is available for comparison; (3) note the heavy emphasis on subjective data, in this case appropriately so because the dependent variable--comfort--could not be measured by objective means.

(2) Ground Emergency Egress Demonstrations

A series of ground emergency egress demonstrations was conducted from the aft troop compartment of a C-5 wooden mock-up . . . in accordance with a contractual requirement to demonstrate the adequacy of escape and survival provisions. This requirement is to evacuate all passengers and crew, totaling 75 personnel, in the event of a ground emergency, in 60 seconds or less with half the exits blocked. Certain aspects of the flight station escape and survival were also examined.

Comment: Exploratory tests are often combined with Resolution and Verification tests.

The mock-up simulated the production configuration. Openings of designed size and location were provided for two forward emergency exits, aft service door and an emergency exit opposite

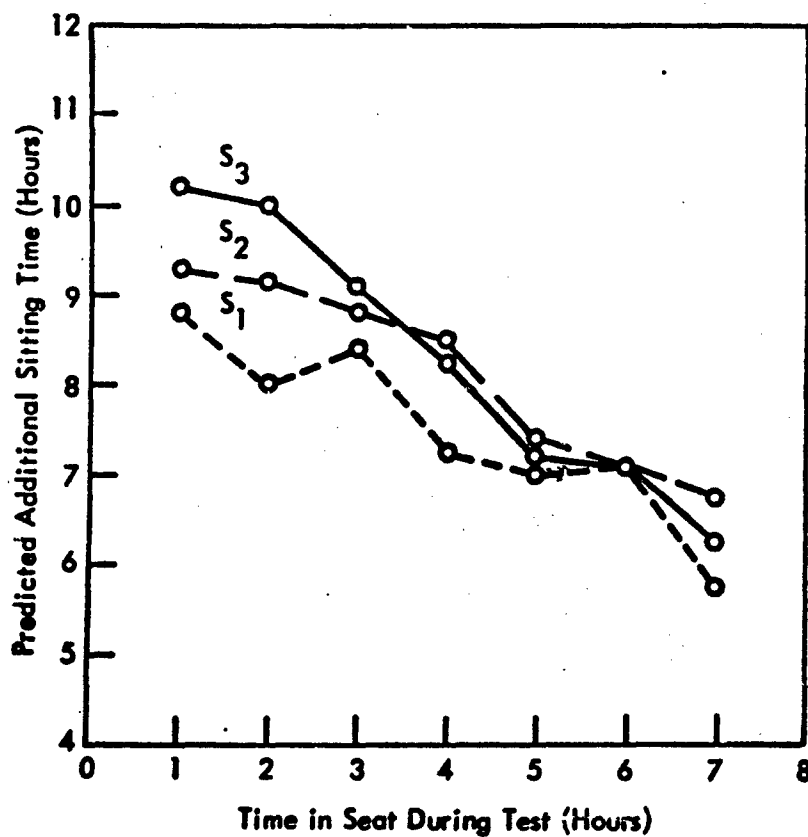


FIGURE 2. AVERAGE HOURLY PREDICTIONS OF ADDITIONAL TIME SUBJECTS ESTIMATED THEY COULD SIT

the service door. An inflatable escape slide was installed at each exit in the deployed and inflated condition. Consistent with the contractual requirement, only two slides were normally used on a given test. (Two tests were run, with four and three slides available respectively, to compute evacuation times under those conditions.) Subjects never knew in advance which exits, or how many, would be blocked.

One demonstration, involving two tests, made use of only a stair-ladder. The stair-ladder, primary means of normal access to the aft troop compartment, could also be used under certain circumstances as a supplement to the emergency egress exits. A ground level mock-up of the flight station/passenger area was also utilized to demonstrate certain features of the ditching survival system and contained a working facsimile of the life raft deployment device to be incorporated in production aircraft.

The subjects were 320 Army personnel divided into four groups of 80 persons each, 75 test subjects and five supervisory officers and non-commissioned officers. An analysis of variance showed no statistically significant differences in age, weight or height among the four groups.

Each new group of subjects received an initial briefing from the test director as to the purpose of the tests and procedures to be followed. Prior to unloading, the subjects donned vests displaying large numbers which were used in later film analysis to plot egress patterns. After all tests except two which involved the stair-ladder, the subjects returned to their seats and completed a questionnaire concerning the test just completed. (This questionnaire was designed to obtain comments and suggestions about needed changes in procedures and equipment, difficulties encountered, clarity of instructions, etc.). After each test the subjects were assigned different numbers so they would not be repeatedly seated near the same exit.

Three types of data were collected. The primary data record was motion picture film. Sixteen mm cameras loaded with color film were positioned at strategic locations. The cameras were synchronized and simultaneously activated when the signal to evacuate was given. Egress time and personnel flow were determined from the motion picture film. Analysis was accomplished using a Photo Optical Data Analyzer which featured a built-in frame counter. Knowing the speed at which the cameras operated (24 frames/sec.) and counting the number of frames for the completion of a test, it was possible to accurately calculate the time required for each test.

As a back-up to the motion picture cameras, trained test observers with stopwatches were positioned at each exit and inside the aft troop compartment to record the stopwatch time after each test. Stopwatch data were used exclusively for the two tests that were run at night.

The third type of data was obtained from questionnaires described previously.

Results of the various tests are shown in Table II. It was concluded that the requirement for evacuation of personnel in 60 seconds or less with half the exits blocked had been met. This conclusion is supported by the results of the tests conducted to satisfy contractual requirements: Tests 2, 5, 7, 10 and 12. With only one exit available, the time required was only 64.67 seconds (Test 11). Should no exits be blocked as a result of the nature of the emergency, the evacuation time was 40.41 seconds (Test 1).

This large-scale quantification of human performance generated data which were beneficial in C-5 mission planning.

b. Operation on the Move: A Feasibility Study, McCommons, R. B. Technical Memorandum 2-72, Human Engineering Laboratories, Aberdeen Proving Ground, Maryland, January 1972.

Tactical units must redeploy frequently to avoid detection, respond to hostile threats and keep pace with normal operations. As a result, many of our weapons and communications systems are mounted on wheeled or tracked vehicles to afford them the mobility needed to support such actions.

Since it was evident the basic limit to system performance during mobile operations would be the capabilities of the human operators, a literature search was done to identify investigations of those capabilities. Finding no significant information on the subject, this investigation was initiated to develop baseline information with which to assess the feasibility of operations on the move. Specifically, the investigation sought to (1) determine how well personnel could perform a variety of psychomotor tasks inside a moving truck-mounted shelter, and (2) identify, if possible, means of enhancing personnel performance.

The basic premise of this investigation was that for operation on the move to be of any value, subjects must be able to perform under realistic conditions; i.e., if either cross-country speed and mobility or operation on the move had to be sacrificed it would be the latter. In keeping with this approach, the subjects were required to do representative command and control tasks while the vehicle traversed courses of varying severity. Vehicle speed was attuned to road surface with the idea of maintaining maximum safe speed.

The subjects (Ss) for the study were six Army enlisted men. These men were selected from a volunteer subject pool temporarily assigned to HFL. The criterion for selection was an ability to type using the home key method. Typing skills of the Ss ranged from 20-60 words per minute.

TABLE II

OVERALL EGRESS TIME AND EGRESS TIME FOR EACH EXIT (In Sec.)

Test No.	Exit				Overall Time
	3L	3R	4	6	
1	40.41	37.58	36.71	35.75	40.41
2 ^o	-	-	41.42	51.87	51.87
3	Down the Stair-Ladder				90.54
4	44.75	47.29	48.12	-	48.12
5 ^o	-	-	48.45	47.50	48.45
6	Up the Stair-Ladder				74.79
7 ^o	59.20	57.92	-	-	59.20
8	All Exits at Night				(48.40)
9	Exits No. 4 and No. 6 at Night				(63.70)
10 ^o	-	58.70	56.87	-	58.70
11	-	-	-	64.67	64.67
12 ^o	44.70	-	-	55.58	55.58

^o Tests conducted to satisfy contractual requirement for demonstration.

() Indicates stopwatch data.

Testing was conducted inside a specially prepared S-141 shelter mounted on an M-35, 6 x 6 cargo truck.

Tables were built along the interior roadside and front walls of the shelter to support equipment and provide workspace. A common variety reclining swivelled office armchair was anchored to the floor at the junction of the "L" formed by the tables and served as the Ss' chair. The test apparatus, consisting of a VRC-12 radio, a desk calculator and a TT-76C teletypewriter, were shock-mounted on the tables to the left and front of the Ss' chair. Behind and to the right of the Ss' chair, a straight-backed armchair was mounted on a raised platform. This chair was provided to allow continuous monitoring of testing by an observer. Footrests were provided for both Ss and the observer.

Three test courses were used in the study. They were selected as being representative of the various terrain features a vehicle might encounter on typical cross-country runs; i.e., presenting vibration having random amplitude and frequency of occurrence. The driver was asked to negotiate the various courses at the maximum safe speed. This was done to provide worst-case vibrational conditions inside the shelter and to assure that cross-country mobility was not being compromised.

Course #1 was a hard-packed, gravel-surfaced road typical of unimproved country roads. This course was a 1.6-mile closed loop having both sharp and sweeping curves; the surface ranged from smooth to rough. Roughness was due to small potholes, washboarding and rutting. The course was driven at 20 miles per hour (mph).

Course #2 was a relatively straight 5.6-mile length of high-crowned macadam road. The generally smooth road surface was characterized by gentle, unevenly spaced depressions (probably caused by frost heaves) which imparted a combination of pitch and roll to the vehicle. The course was driven at 35-40 mph.

Course #3 was the 1.8-mile long Cross-country #2 section of the Perryman Test Area at Aberdeen Proving Ground. This course had sweeping curves and the hard-packed gravel surface varied from smooth to rough. Sections of the course contained large, unevenly spaced potholes and deep ruts. The course was driven at 10 mph.

Of the three courses, Course #1 was considered the least severe and Course #3 the most difficult.

The Ss were required to perform three different tasks: (1) operating a key entry device (the desk calculator), (2) teletypewriting and (3) radio tuning.

These tasks were chosen as being representative of command and control functions personnel might perform in command post or fire control center operations.

First, the Ss were thoroughly acquainted with the purpose of the study and the testing procedures that would be followed throughout the experiment. Then, each S was individually trained to use the test apparatus (i.e., desk calculator, radio and teletypewriter) and subsequently given practice trials doing the experimental tasks.

The operations of the desk calculator and radio were relatively simple. Therefore, the Ss were not held to a strict training schedule when being instructed in their use. Training was considered complete when, in the experimenter's opinion, the Ss were performing the required tasks in a facile manner.

Time constraints made it impractical to attempt training the Ss to peak performance levels on the teletypewriting tasks. Therefore, training and practice with the teletypewriter was continued until each S had reached a performance plateau. This plateau was considered to have been attained when the S exhibited consistent performance on three consecutive practice trials.

After training was completed, the actual testing began. Testing for each S consisted of three sessions. In general terms, a test session consisted of having the S perform each of the three selected tasks under three operational conditions: static (vehicle not moving), Course #1 and Course #2 (Course #3 was run separately and is discussed later in the report.) The sessions normally lasted 2-3 hours and were conducted at least one day apart.

Typically, the S was met at his barracks and ushered into the shelter. After the S's chair and copyholder was positioned to his liking, the S was briefed on the test procedures. Then, depending on the condition to be run first, either testing began or the vehicle was driven to the appropriate course. If a road course was scheduled first, the S rode in the truck cab to avoid undue fatigue.

Before starting any road course, the test driver and the observer established communication via walkie-talkie. Assuming all was well, the driver was instructed to accelerate the vehicle to the speed selected for that course. When the observer was notified the desired speed had been reached, testing of the S was begun. If, for any reason, the predetermined vehicle speed could not be maintained, testing was halted. Testing was resumed as soon as conditions permitted.

Depending on the task scheduled first, the S was given a worksheet instructing him to do a set of arithmetic problems using the desk calculator, type prepared text, or set up the radio for push-button channel selection.

After the first task was finished, the S returned the worksheet to the observer and received new instructions. When all the tasks were completed, the driver was notified and the vehicle was stopped. The S was then allowed a rest period. This same general procedure was followed until the S had performed all the tasks under all the conditions.

After all the trials on Courses #1 and #2 had been completed, a cursory examination of the data revealed that the capabilities of the Ss had not been fully explored. Therefore, the Ss were asked to participate in a final, and more taxing, phase of the study. It was at this time that Course #3 was run. Time constraints allowed only two runs of this course for each of the six Ss. The general testing procedure was the same.

Two measures of S's performance, task completion time and error frequency, were obtained for each test condition. Task completion times were measured with a restartable stopwatch and recorded by the observer. Error frequency on the key entry and teletypewriting tasks was determined from hard records; i.e., completed worksheets and perforated-tape output from the teletypewriter. Errors in radio tuning were noted visually and recorded by the observer. The data were analyzed with a series of uncorrelated t tests.

In general, the results of the study were threefold. First, and probably most important, the Ss were able to effectively perform the tasks asked of them under conditions that, subjectively, ranged from mild to severe. Second, the primary effect of vibrational severity on the Ss' performance was to increase the time needed to complete each task; i.e., time varied directly as a function of course severity. Third, although task completion times increased, the quality of the finished product was comparable under all the conditions tested. This indicates, that, although the vibrational environment may have caused more errors to be made, the errors were caught and corrected.

2. Resolution Tests

a. Evaluation of Launch Vehicle Assault Motion Effects on Personnel Performance, Stinson, W. J., Navy Personnel Research and Development Center, San Diego, California, June 1977.

This is another example of a Resolution test combined with an exploratory one: to determine whether personnel could withstand the motion effects of a proposed new vehicle by comparison with the effects experienced with an already available vehicle (the LVTP-7).

Future amphibious landings must be initiated at greater distance offshore for protection against long-range weapons fire during low-speed launching and form-up phases. An advanced Landing Vehicle Assault (LVA) platform is being developed by the Naval Sea Systems Command . . . for Marine Corps use. Upon Launching from over-the-horizon amphibious force ships, the LVA will transport troops at relatively high speed (25 mph or more) to beach or inland combat positions. The LVA will eventually replace the existing, much slower LVTP-7.

Candidate vehicle configurations under consideration for this amphibious mission include planing hull and air cushion types. An important factor which must be taken into account in the selection process concerns the effects of high-speed motion on troop performance. Past experience has indicated that debilitating seasickness (kinetosis) is commonly experienced by troops during amphibious landings. The potential for troop performance decrement associated with motion effects would normally be expected to increase as LVA speed is increased by a factor of three or more over the LVTP-7.

Current LVA work at the Navy Personnel Research and Development Center is limited to comparative evaluation of experimental planing hull and LVTP-7 motion effects on performance.

A full scale hydrodynamic vehicle (FSHV) served as an experimental LVA planing hull for use in coastal trials (Camp Pendleton area). The FSHV was similar in size, weight, and speed characteristics to the future LVA. However, troop carrying capacity was limited to permit accommodation of commercially available engines, test instrumentation, and underway observers. Although troop compartment capacity was limited to nine men, the space provided per man was equivalent to that available in the LVTP-7.

Experimental hull dimensions were largely predetermined by LVA mission-related specifications involving the eventual number of troops to be transported (about 25 each vehicle) and the need for size compatibility with amphibious ships well deck storage space Within these limitations, the hull shape had to contribute to attainment of speed objectives while serving as a reasonably stable platform in moderate sea states and retaining full capability as an armored personnel carrier upon landing. The vehicle must be able to operate at maximum speed in sea state 2, reduced speed in sea state 3, and survive in sea state 4. The experimental LVA/FSHV configuration is illustrated in Figure 1 (10).

Small craft such as the LVA/FSHV are particularly vulnerable to wave-induced motions. The LVA/FSHV hull shape . . . was designed to reduce motion effects to reasonable levels.

The purpose of this study was to demonstrate that the performance of troops landing aboard the high-speed LVA/FSHV after 1 hour open ocean transit was at least as good as that of troops landing aboard the existing LVTP-7 after 30 minutes transit under similar sea state conditions. The variation in transit time reflected the typical mission scenario applicable to each vehicle. The requirement for equal or better performance by FSHV troops even though water-borne exposure time is doubled obviously placed a heavy burden upon the experimental vehicle in demonstrating acceptable ride quality and habitability characteristics.

As part of the FSHV design process, special attention was directed toward controlling potential habitability problems (noise, ventilation, and air quality). Design provisions were believed to be adequate for the purpose and in compliance with MIL-STD-1472B. Measurements were made aboard both the FSHV and LVTP-7 before and during landing trials to verify environmental adequacy. However, there was no attempt to vary such factors as ventilation, temperature, and noise for evaluation of possible effects on troop performance. Ambient conditions will naturally change from time-to-time in conjunction with sea state and weather conditions.

The principal concern of this study involved comparison of different LVTP-7 and FSHV high speed motion effects on troop performance. There was no attempt to precisely simulate stationary pre-launch (well deck) and low-speed group form-up operations (other aspects of the LVA mission). The effects of crowded confinement over a substantial period of time (possibly 1 hour or more) prior to high speed transit should be similar for each vehicle, except as affected by differences in habitability conditions (ventilation, temperature, air quality, and noise). While the effects of different habitability conditions during stationary and low speed operational phases were of some general interest, practical time and financial constraints required limitation of the scope of this study to examination of high speed motion effects.

If test objectives were successfully achieved, the FSHV would demonstrate that a planing hull configuration with properly designed habitability features can deliver troops with performance capabilities at least as good as those of troops transported aboard the existing LVTP-7. The LVTP-7 has been in wide-spread use for several years and the performance level of assigned troops has apparently been acceptable. This performance level has therefore been adopted as a yardstick against which the FSHV must compare favorably.

Test subjects participated in orientation training and practice in performing test tasks prior to commencement of landing trials. Pretesting of subjects was accomplished to establish individual baseline scores. Subsequent performance during landing trials was then compared to evaluate the effects of transit conditions.

The tasks selected for use in performance testing were intended to be representative of Marine activities normally associated with beach landing operations. Combatant troops must be able to move quickly and fire accurately after transit through coastal waters. Time expended and problems encountered in traversing an obstacle course upon landing were recorded for each test subject. Accuracy in a rifle-firing exercise following the obstacle run served as a principal criterion measure.

Test facilities had to be located in the immediate vicinity of the FSHV and LVTP-7 landing area. Transportation to a remote firing area was not permissible inasmuch as the effects of waterborne transit would be changed during further land travel. Performance testing had to commence immediately upon landing and be completed within one hour.

Considerable resources were expended in conducting each trial run. The time allocated to each cycle had, therefore, to be effectively utilized in collecting data directly relevant in evaluating LVTP-7 and FSHV motion effects. Other data of secondary concern could be collected on a lower priority "not to interfere" basis.

Prior to getting underway, both vehicles were stationed at the Del Mar Boat Basin, Camp Pendleton. The FSHV was moored at a dock for loading while the LVTP-7 was loaded ashore. Loading and deployment at low speed (about 5 mph) through the Oceanside harbor channel involved about 15 minutes for each vehicle. Subsequent open ocean transit then involved 30 minutes for the LVTP-7 and 1 hour for the FSHV. Return through the channel and unloading at the landing site again involved 15 minutes for each vehicle.

Special precautions were taken to avoid conditions which would endanger the safety of test participants. Run speeds for the FSHV started at 15 mph and gradually increased to 30 mph, subject to demonstration of satisfactory seakeeping characteristics at each 5 mph increment. Run speeds above 30 mph (possibly to 40 mph) were optional. Open ocean operations were conducted within a maximum range of 10 miles from shore. High speed landing operations were typically conducted in sea state 2. This mode is of principal concern in evaluating troop performance. Reduced speed operations were conducted occasionally in sea state 3. This mode was of low priority concern (optional) for investigation within project time constraints. Trial runs were arranged to assure that comparable sea state conditions were experienced by the FSHV and LVTP-7. Each vehicle was scheduled to complete a run in the morning and afternoon. Thus, 12 test days would be

required to complete 24 sea state 2 runs for each vehicle. It should be noted that the 12 days of test operations could occur over a period of 2 or 3 months due to interruptions involving inclement weather/sea conditions, vehicle/instrumentation malfunctions, vehicle technical tests, personnel time off, etc.

Inasmuch as this effort was primarily concerned with evaluating waterborne motion effects on combatant troop performance, preferred test subjects were experienced Marine infantrymen (at least 1 year of service). As a safety precaution, subjects had to possess Class 2 swimmer qualifications. Two groups of subjects were assigned with a Squad Leader (NCO) and 8 Riflemen in each group.

Subjects were equipped with marching back-packs (light weight), flak jackets, life preservers (Mae West type), and helmets. As a safety precaution, rifles were not carried inside the vehicles but were issued and controlled at the rifle range.

The performance of personnel assigned to vehicle driver (or alternate driver) duties was evaluated separately from that of embarked rifle squad troops. Driver performance was monitored in steering predetermined course patterns, requiring heading changes from time to time upon command. Time to achieve ordered headings was recorded. Additionally, driver/alternate driver personnel participated in rifle firing on a limited basis (at the end of each two vehicle runs) to provide an indication of eye-hand coordination capabilities.

Six test supervisors/observers were assigned. These personnel monitored the overall test effort and assisted at various positions as needed.

A Marine observer/crew member was assigned to the Troop Commander duty station in each vehicle, with access to VHF radio communications equipment. These personnel secured the vehicles upon landing and assured that clean-up/maintenance work was satisfactorily performed in preparation for subsequent operations.

Industrial hygienist personnel monitored environmental conditions aboard the FSHV and LVTP-7 during landing One observer assisted in monitoring environmental conditions aboard the vehicle and a second observer monitored driver/alternate driver performance.

Several observers recorded troop task performance as rapidly as possible after landing (prior to wear-off of motion effects). Three observers were at the obstacle course to record the speed and problems of runners in two lanes, starting 15 seconds apart. One observer was at the rifle range to prepare targets/ammunition, check firing line readiness, and order commencement/termination of firing.

Evaluation tasks consisted of running an obstacle course and rifle firing. (The ideal measurement situation would have been to mount an actual combat exercise but this was impossible because of cost.) The obstacle course started at a point on the beach within about 1,000 feet of the FSHV docking platform and LVTP-7 landing area. The course covered a distance of 300 feet. Two parallel running lanes were provided, with five pairs of obstacles located at intervals along the lanes. The obstacles included a 5 foot vertical wall, inclined balance logs, cargo net ladder, horizontal balance beam, and staggered tires.

These tasks were intended to test balance, agility, and coordination (all presumably required in amphibious combat). Time to negotiate the obstacle course was recorded for each test subject. Upon completing the obstacle run, subjects proceeded without delay to the rifle range (about 1,200 feet away).

The type of weapon used had to be compatible with provisions which can be made for safe facilities in close proximity to the vehicle landing site. Precision air rifles with verified accuracy were used to fire pellets at bullseye targets. The type of rifle used for test purposes had no significant impact on evaluation of motion effects since performance was compared against baseline scores established with the same type rifle.

Upon completion of rifle range operations, subjects were given a short rest (5 minutes) prior to returning to the obstacle course for repetition of test tasks. Results of the repetition cycle provided an indication of changes in performance as motion effects wore off and reached an insignificant level.

A repeated measures design was used in the collection of performance data. In this design, each subject was required to perform the selected tasks several times to establish baseline reference scores prior to commencement of landing trials. This permitted measurement of variance in individual performance since the tests were repeated with each subject serving as his own control. Under this procedure, the use of separate control group subjects was eliminated and inter-subject variance was not a problem.

Upon completion of all trial runs and recovery from motion effects, subjects performed a final series of test tasks to establish posttest scores. This indicated any natural improvement in proficiency as the result of "practice" in repetitively performing tasks during the test period. This "practice" effect had to be taken into account in evaluating test results. The pretest baseline scores of individual subjects were compared against subsequent performance during landing and posttest trials. In addition, the overall performance level of the two squads was evaluated in relation to each other.

The sequence of FSHV and LVTP-7 runs was rotated on successive days so that morning and afternoon landings were balanced between the two vehicles.

Preliminary results of the study showed no significant differences in personnel performance following transits in the FSHV and the LVTP-7. (A more detailed report of results is not available because the data are still being analyzed.) This represented a desirable conclusion to the study because significant differences would have indicated (a) either that the FSHV was inferior to the LVTP-7 (highly undesirable) or (b) that FSHV was superior to LVTP-7 (practically speaking, very difficult to demonstrate because of the increased stress of the FSHV operating conditions).

3. Verification Tests: Operational System Test

The following summaries are taken (with some editing) from Askren, W. B. and Newton, R. R., Review and Analysis of Personnel Subsystem Test and Evaluation Literature (Report AFHRL TR-68-7), Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, Ohio, January 1969.

a. Titan II Inertial Guidance System

A. C. Electronics Division, General Motors Corporation, Inertial Guidance System, Weapon System 107A-2, Category II Personnel Subsystem Test and Evaluation (PSTE) and Maintenance, Logistics, Reliability and Readiness (MLRR) Test, and Evaluation, Document 64-197, Milwaukee, Wisconsin, June 1964.

a. Scope and Relation to (PS) Elements

(1) This report covers the results of the Category II PSTE/MLRR test program conducted by A. C. Spark Plug Division of General Motors Corporation on the Titan II IGS.

(2) Personnel Subsystem Test and Evaluation covers that part of weapon system testing that involves human engineering, personnel, training, and the validation and verification of technical publications.

(3) The PS, as defined in this report, is a composite of all performance elements in a system assigned to man (both operational and maintenance) and the means for implementing such elements, including not only the assigned men but also any physical products provided to support man's performance, such as equipment, tools, facilities, spares, and technical orders. Equipment and tools used by man to perform a function

is considered to be an element of the personnel subsystem, but equipment operated or maintained by man is an object of the personnel subsystem, and not a part thereof. In the Titan IGS, the personnel subsystem is a part of the maintenance subsystem.

b. (PS) Test Objectives

Specific test objectives were listed for personnel performance, safety, technical data, maintenance, logistics, reliability, readiness, and weapon system capability. Examples of these test objectives were:

(1) Personnel Performance (12 objectives).

(a) To determine whether qualified military personnel can effectively prepare, operate, and maintain the weapon system using only authorized equipment procedures.

(b) To evaluate the effect of observed human engineering, procedure, and personnel deficiencies on the weapon system incommission rate.

(2) Safety (7 objectives).

(a) To determine whether propellant transfer can be safely operated and maintained.

(3) Technical Data (7 objectives).

(a) To evaluate procedures altered because of equipment changes.

(4) Maintenance (6 objectives).

(a) To determine whether time required to perform scheduled maintenance tasks is within prescribed Technical Order (T.O.) limits.

c. Data Requirements and PS Test Criteria

(1) Attitude data of technicians.

(2) Performance deviations from T.O.'s.

(3) Time to perform operations.

(4) Test criteria included "adequacy" (availability and system accuracy) and "efficiency" (expenditure per unit of output).

d. Data Collection Methodology

(1) Use of location logs to determine availability.

(2) Training equipment was used on a noninterference basis to evaluate unscheduled maintenance activities. Six Ballistic Missile Inertial Guidance Technicians (AFSC 312X2F) were assigned as test subjects for this program. Under observation and by design they performed troubleshooting tasks on the Inertial Guidance System (IGS) trainers, which allowed validation of Trouble Analysis Diagrams and procedures.

(3) An evaluation packet was constructed for each scheduled test session. Additional data collection forms were provided for collecting data during unscheduled maintenance activities.

(4) Two methods of collecting data were implemented in the field. The major set of data were provided by the Observer/Evaluator (O/E) and were based on his observations of the activities performed by Air Force technicians. When the O/E was not available during unscheduled activities, A. C. Spark Plug engineers and technicians, on location, collected basic data.

(5) Data collection forms and methods were designed to maximize technical information and minimize clerical information.

(6) Data collection method was oriented towards collecting true performance data without affecting the performance, and collecting additional data subsequent to the performance. Data collection forms contained reminders and cues for the O/E. Time data were obtained by having O/E's document step numbers and running time for all problems noted.

(7) The method was oriented towards evaluating all of the means used to implement a given function, not oriented specifically toward performance of personnel. Thus, problems in the operations of an equipment item were considered to be problems in personnel performances.

(8) The data collection method was also designed to enable the analysts to completely reconstruct the test session on paper.

(9) Copies of the maintenance data, form, special unit record, and location log were presented in the report.

(10) Most of the functions critical to the operation of the system were evaluated a sufficient number of times to obtain a good representation of those functions.

(11) Three personnel performance detectors were used: one evaluator was from the 3901st Squadron; second evaluator was from Quality Control and Evaluation; and the third evaluator was the supervisor. Data suggested performance was different under each type of observer.

f. Reducing and Analyzing Data

(1) Over-reaction to personnel subsystem problems was avoided by relating them to system effect.

(2) Personnel performing analyses were experienced in the design and development of the system.

(3) Three analyses were performed concurrently on the data received from the field;

(a) Personnel performances were analyzed to estimate efficiency, adequacy, and reliability.

(b) Investigations of deficiencies were reported on Individual Summary Forms. (Comment: Problem solution)

(c) Estimations were made of systems "use" reliability.

(4) Analysis was performed on four types of personnel performance errors: Type 1 (rejecting a good unit), Type 2 (accepting a bad unit), Type t (increasing performance time), and Type d (incurring damage).

(5) The impact of personnel performance problems was determined by using the Maintenance Subsystem (MSS) Model which related these problems to overall systems measures such as availability and levels of spares.

(6) Analysis of personnel performance data considered the kind of evaluator who furnished the information, as major differences in performances were noted when technicians were observed by different kinds of raters, e.g., 3901st or O/E's. (Comment: This discrepancy is understandable, but unnerving.)

g. Significant Test Results

(1) The personnel subsystem was a major source of system inefficiency. Primary causes were incompatibilities in the assignment of skill levels, the tendency to treat the personnel as automatons, and a lack of a systematic means of evaluating and maintaining the personnel at the level required by the system. Skill levels assigned were too high for operations required.

(2) A major omission in the personnel subsystem appeared to be the lack of evaluation and practice of emergency procedures at the launch sites.

(3) Significant differences were noted between operations at Vandenberg and at operational sites.

(4) (a) Unrealistic policy of "strict" adherence to T.O.'s caused morale problems. (b) Two men were used for many one-man tasks. (c) There was a lack of sufficient training and data to operate effectively when a malfunction was not adequately covered by T.O.

(5) The individual test objectives were evaluated as to whether met: "yes," "no," or "partially."

(6) Based on data in evaluation packets, no difference between the bases was found in personnel performance.

(7) Percent contribution to downtime due to inefficient personnel performance was 11.5%.

(8) Many deviations from T.O.'s, by high skill personnel, were found.

(9) Many errors in T.O.'s were noted.

h. Communicating and Using Test Results

(1) Details of the data collection during the Category II PSTE/MLRR program were reported in monthly Detailed Analysis Reports.

(2) Correction of the problems identified would result in a considerable cost savings due to a reduction of skill levels for a large percentage of the performance elements in the system and a reduction of manpower.

(3) The results of the analysis were reviewed periodically to determine the trade off analysis to be performed, the MSS Model variables to measure, and the MSS Model variables to study. Recommendations were made to engineering management on the trade off which should be performed.

(4) Exercising the MSS Model on the International Business Machine 7090 computer enabled A. C. Spark Plug to validate past decisions regarding spares provisioning, personnel Maintenance Ground Equipment utilization, and maintenance configurations.

. . .

k. Other Problems

(1) Major problems encountered during the initial phases of the Category II PSTE/MLRR program were the lack of representative conditions at Vandenberg Air Force Base and the lack of opportunity to collect personnel performance data. These problems were resolved, with the cooperation of the Ballistic Systems Division and the Strategic Air Command, by transferring the data collection program to the three operational bases.

(2) Another major problem encountered during the program was the lack of data required to evaluate the condition of the system and the adequacy of the time interval between scheduled checks.

(3) An output of the Sheppard program of evaluating troubleshooting problems on trainers was the additional training provided Strategic Air Command (SAC) guidance personnel. The test subjects were then able to inaugurate effective on-the-job training for other SAC personnel at their respective bases.

b. System 412 L

Adams, J. A. and McAbee, W. H., Lt. Col., USAF, A Program for Evaluation of Human Factors in Category II Testing of Air Weapons Control System 412 L (Phase II Configuration), PGN Document-62-1, Deputy for Bioastronautics Air Proving Ground Center, Eglin Air Force Base, Florida, May 1962.

(Comment: It is not known whether the actual test was conducted fully in accordance with the following test plan.)

Human factors variables will be weighed insofar as they influence criteria of success in achieving combat goals of the Air Force. (Comment: The personnel subsystem is important only insofar as it influences system success.) These variables should be examined for three basic configurations of the system: (1) a normal system where all automatic subsystems are fully functioning and the personnel subsystem is playing its assigned role, (2) a normal system that is degraded by electronic counter-measures, and (3) a fully degraded system whose operation is essentially manual. System degradation is considered a realistic expectation for combat uses of AWCS 412 L, and the personnel subsystem increasingly contributes to the accomplishment of system goals as degradation increases. If the success criteria are not met for one or more of these three system configurations, human factors variables related to human engineering, training, organizational structure, and maintenance will be studied to diagnose the cause of the system's deficiency. Biomedical determinants of personnel safety, and variables related to the handling and assembly of components of the mobile system, will also be studied.

a. Scope and Relation to PS Elements

(1) The emphasis in this plan was on human performance variables influencing the effectiveness of the system as a whole.

(2) The emphasis in measurement was on indices of system performance that reflected the simultaneous actions of all men and machine elements as they bear on the accomplishment of Air Force goals.

(3) According to the report, one of the most serious errors that can be made in the testing of a military weapon system is to limit the evaluation to normal system uses, i.e., noncombat training uses. Test programs must provide for exercises with inputs that closely simulate the tactics and equipment of a potential enemy, where possible.

b. PS Test Objectives

(1) The personnel subsystem tests and evaluations were undertaken to determine how well the human components of a weapon system performed their assigned functions and to identify those changes in procedure and/or equipment design that could increase the effectiveness of the total system.

(2) The purpose of evaluating human factors in Category I, II, and III testing was to determine if the system design, with respect to the men who use it, was adequate to accomplish the combat mission assigned to the system.

c. Data Requirements and PS Test Criteria

The following are criteria of system success for use in diagnosing shortcomings in the personnel subsystem, although some changes in these criteria may be needed to insure compatibility with the criteria used by the test team for other purposes:

(1) Number of Penetrations into the Defense Area- The number of enemy weapons that penetrate this area would be inversely related to the system effectiveness.

(2) Kill Range of Intercepted Targets- This measure assumed that the more effective system would destroy targets at longer ranges.

(3) Number of Kills or Probability of Kills- This factor would simply be a matter of recording the number of kills or probability of kills, as determined by an accepted criterion.

(4) Number of Assignments of Weapons to Targets- This measure assumed that the most efficient system would destroy targets with the least number of weapons.

(5) AWCS 412 L was concerned not only with destruction of enemy targets, but also with the speedy and safe recovery of aircraft. A fundamental measure for aircraft recovery would be holding time. There may have been delays in landing procedures to the extent that the system was weakly designed for recovery operations.

d. Data Collection Methodology

The Playback Camera that regularly photographed the system's status displays was a candidate as a basic instrument for measuring system performance. The evaluation of the film frames

should have allowed tabulation of the number of penetrations into a defense area, kill range of intercepted targets, number of kills, number of assignments of weapons to targets, and holding time.

. . . .

h. Communicating and Using Test Results

The test results were to be used to diagnose shortcomings in the system. The locus of these shortcomings would then be determined through further testing.

c. Atlas "D"

Air Force, Golden Ram Personnel Subsystem Final Report, Atlas "D" Series, 8 September 1961, contract AF 04(647)-619.

a. Scope and Relation to PS Elements

This program is primarily concerned with the following areas:

(1) Equipment Design: Human design considerations relative to maintainability, operability, etc.

(2) Technical Data: Technical publications used by personnel to operate and maintain the system.

(3) Job Environment: Work conditions which affect personnel performance.

(4) Personnel Selection and Manning: The adequacy of the number and the quality of Air Force Specialty Codes (AFSC's) used on the job.

(5) Training: The adequacy of training received and the means by which it was imparted to personnel.

(6) Organization Control Procedures: Procedures used to control personnel in the operation, maintenance, and management of the hardware.

(7) Technical Representatives: The utilization of Contractor Technical Services Personnel.

b. PS Test Objectives

The objective of the Personnel Subsystem's portion of Project "Golden Ram" were to determine whether the Atlas Series "D" weapon system was capable of being operated,

controlled, and maintained by qualified Air Force personnel. (Comment: Note how the Exploratory aspect blends into the Resolution. This is because no quantitative personnel standards existed for Atlas "D.") Data collected assessed the satisfactory or unsatisfactory nature of the Personnel Subsystems operating within the context of the weapon system mission and was used to initiate further investigation or corrective action.

. . . .

d. Data Collection Methodology

(1) Maintenance and operating personnel were observed as they performed assigned tasks using "Golden Ram" checklists. Observer personnel were assigned on a one-for-one basis, for each individual observer except in certain instances. Due to confined space, shortage of qualified observers, etc., the observer to operator ratio decreased. Each observer had a copy of the checklist so they could detect any deviation from the conditions stated on the checklist. Deviation from the operational checklist and difficulties the job participants encountered due to insufficient tools, poor lighting, equipment design, safety factors, etc., were recorded on the observer's Personnel Performance checklist. Observer personnel did not converse with the individuals being observed during task performances.

(2) Subsequent to completion of the operating tasks, each observer interviewed the individual observed. A Post-test Interview Form, consisting of twenty-seven (27) operational type questions and twenty-three (23) troubleshooting questions, was used. Spontaneous questions were introduced by the interviewer where applicable. Operational type questions were asked after each job performance. Troubleshooting questions were asked only after there had been an unscheduled maintenance operation or deviation from the checklist.

(3) Following conclusion of the post-test interview, the observer reviewed the data recorded on Post-test Interview Forms, Personnel Performance Checklists, and initiated Deviation/Difficulty Reports (D/D's) on this information.

(4) Data was collected in as near an operational environment as possible.

(5) Work was performed without interference from the observer force.

(6) Aptitude tests and background interviews were administered to thirty-two (32) military personnel and forty-four (44) contractor technicians. These tests were administered and the results compared in an attempt to determine if the Atlas "D" series missile was too complex for the Air Force technicians. The areas encompassed by the test were: verbal comprehension, numerical ability, visual pursuit, visual

speed and accuracy, space visualization, numerical reasoning, word fluency, manual speed and accuracy, and symbolic reasoning.

...

f. Reducing and Analyzing Data

(1) Deviation/Difficulties, whether discovered through interview or observation, were recorded. The D/D Forms, along with the post-test interview sheets and Personnel Performance Checklists, were forwarded to, and processed by the Data Analysis Section. Processing included numbering each D/D Form and entering it on a master control log providing overall control of total number and type. Each D/D was then reviewed for technical validity and analyzed for system degradation implications, as appropriate.

(2) When an operation was completed, all D/D's written were categorized and assembled into a Summary Analysis Report. This report described the operation and the Deviation/Difficulties recorded. Implications and problems were defined and an overall analysis of the D/D's accomplished.

g. Significant Test Results

(1) Air Force personnel compared quite favorably to the contractor personnel on the aptitude tests.

(2) Qualified Air Force personnel could operate, control, and maintain the Atlas Series "D" weapon system satisfactorily.

(3) Results of the Personnel Subsystems tests disclosed deficiencies in the Atlas "D" series weapon system which could have prevented the weapon system from responding satisfactorily, had no corrective action been taken and in some areas, further action was imperative. Deviation/Difficulty statistics revealed potential trouble areas as follows:

(a) Technical Data- This area presented the most D/D's. Inadequate or incorrect technical data degraded the weapon system. Evaluation was not as effective as desired due to continued technical data correction.

(b) Training- Deficiencies in Individual Operational Readiness, and On-the-Job Training appear to have existed; however, specifics could not be identified due primarily to excessive time lag between Individual Training, Operational Readiness Training (ORT), and on-site AFSC performance. Overall Training, as a package, however, is sufficient.

(c) Equipment Design- Relatively few D/D's found in this area.

(d) Personnel Selection and Manning- No problem was found in this area.

(e) Job Environment- Only minor deficiencies found in this area except in communications. Other than communications, the overall environmental situation seemed to be satisfactory.

(f) Technical Representation- No D/D's were written against the qualifications of technical representatives.

(g) Communications Discipline:

A number of D/D's were generated as a result of lack of communications discipline.

Standardized communications techniques and procedures were not utilized in many instances.

The crew chief was overloaded with administrative details.

(h) Nonessential Visitors- During Project "Golden Ram," an excessive number of noncrew personnel were present.

(i) General Electric MOD III (Modification Three) Guidance- Personnel Subsystems Observations of this area were limited.

(j) Weapon System Personnel Subsystem Supervision and Control:

Logistics presented continued problems. Nonavailability of tools, test equipment, and parts were continually hampering operations.

Crew Discipline and Supervision- Personnel Subsystem observers were frequently reporting failure of an AFSC to sequentially follow the checklist.

Safety- Safety deficiencies were predominantly attributed to carelessness.

Training- discussed under "b" above.

h. Communicating and Using Test Results

(1) Recommended corrective action was introduced and inserted on the D/D's form. These D/D's were then reviewed by a board of cognizant military and contractor personnel. If accepted, the D/D's were reproduced in final form and forwarded to the appropriate action agency. On return from the action agency, the corrective action was reviewed and either closed out or sent to the Test Working Group for further action.

(2) When an operation was completed, all D/D's written were categorized and assembled into a Summary Analysis Report. This report described the operation and Deviation/Difficulties recorded. Implications and problems were defined and overall analysis of the D/D's was accomplished.

. . . .

k. Other Problems

Data limitations were imposed by these factors:

(1) Technical limitations.

- (a) Inadequate technical data.
- (b) Incorrect technical data.
- (c) Shortage of tools, test, and safety equipment.
- (d) Limited number of samplings.
- (e) Lack of true operational environment.

(2) Administrative and Personnel Limitations.

- (a) Contractor observers were withdrawn in early stages.
- (b) Constant rotation of observers.

(3) Individual and Operational Readiness Training Deficiencies.

(a) Except in four instances, it was impossible for Personnel Subsystems to properly evaluate either Individual or Operational Readiness Training (IT/ORT) due to an excessive time lag from graduation of IT/ORT and actual beginning of "Golden Ram" on the date the AFSC was assigned operational site duties.

Comments: The following characteristics of the OST should be noted.

1. Test objectives cover all discernable aspects of personnel performance.
2. Data collection methodology makes extensive use of subjective reporting methods and large numbers of observers.
3. Determination of the impact of personnel performance on system effectiveness is a major concern. Only those deviations/difficulties that pose substantial danger to system effectiveness are selected for remedy.

4. Experimental Field Studies

a. Facilities Maintenance Demonstration Study (NPRDC TR 76-29), Schwartz, M. A. Navy Personnel Research and Development Center, San Diego, California, January 1976.

FM*, as currently performed, at-sea and in-port, by ship's force, requires a considerable expenditure of man-hours and material resources. It is estimated that in excess of 1380 man-hours per week, or approximately 27 man-weeks (i.e., 27 men, working full time) is spent on FM aboard an FF 1052 class ship. This represents approximately 11 percent of the man-hours worked by the total enlisted crew.

The objective of the study was to devise, demonstrate, and evaluate methods for reducing shipboard FM man-hour expenditure while improving readiness and condition of ships.

Three basic aspects of shipboard FM were attacked simultaneously: (1) manpower organization and information management, (2) training and technical information, and (3) equipment, materials, and environmental improvements.

MANPOWER ORGANIZATION AND INFORMATION MANAGEMENT

The following three concepts formed the basis for innovations in this area:

1. One specialist team could perform all FM more quickly and efficiently than it is performed using the current personnel assignment methods.
2. Individual FM tasks could be consolidated and grouped according to job type and space or surface characteristics. The redefined job could then be done more efficiently by members of the FM team.
3. An information management and task scheduling system, similar to the existing Planned Maintenance System, could be developed and used to ensure systematic accomplishment of the FM work.

A prototype ship's instruction was prepared which provided information regarding the establishment of the specialist, eight-man, FM team.

A prototype management information and task scheduling system was prepared on the basis of space and FM task analyses. The elements of the system included:

1. A Job Information Card (JIC) for each consolidated set of tasks and spaces.

*FM is cleaning the ship.

2. A master schedule plan for distribution of JICs to billets.

3. Instructions for using the system.

The system was to operate as follows:

1. The work center supervisor, at the beginning of each week, was to determine which specific JICs were to be used.

2. The supervisor would distribute groups of JICs to the team billets. The individual team member receiving a set of JICs would then know exactly which tasks he had to perform.

3. After completing the task shown on the JIC, the FM team member would record, on the JIC, the data required and would return the completed form to the supervisor.

4. The JICs could then be used to update the master schedule of FM tasks.

TRAINING AND TECHNICAL INFORMATION SUPPORT

A prototype FM training program was developed for shipboard use, which consists of 13 audiovisual modules dealing with various aspects of FM.

Each training module consists of a set of 35mm slides and a magnetic sound tape recording. Standard 35mm slide projectors and tape players or synchronized projector/sound units were used to present the modules.

Most of the modules show, in step-by-step fashion, how to accomplish specific shipboard FM tasks. The rest deal with general training, such as safety.

A variety of recent FM equipment and materials was examined to determine its potential for labor and cost savings, safety, and effectiveness.

A master implementation and test plan was devised. Extensive coordination between contractors and the fleet was required to ensure timely and proper installation.

STATEMENT OF HYPOTHESES

In accordance with the objective and goals of the present study, the following hypotheses were established:

1. The implementation of the aforementioned innovations will result in a reduction of FM man-hours.

2. Appearance and cleanliness of the spaces maintained by the FM team will be judged to be adequate or improved.

3. FM team members will demonstrate that their knowledge of FM requirements, techniques, materials, and procedures has increased.

TEST VARIABLES AND MEASURES

The independent variable used in this study was the entire set of innovations. Two basic types of comparisons were planned: (1) conditions "before" vs. "after" on the test ship, and (2) test ship condition vs. control ship(s) condition.

The dependent variables were: (1) FM man-hours, (2) cleanliness and appearance of shipboard spaces, (3) FM skill and knowledge, and (4) attitude and motivation.

The measures of FM man-hours were:

1. Estimates of FM task times on control ships.
2. Estimates of FM task times from documented sources.
3. Actual recordings of FM task times from completed JICs.
4. Comparisons (on the test ship) between subjectively estimated job times before and after innovations were installed.

Measures of appearance and cleanliness of spaces consisted of (1) completed inspection rating forms using subjective scales and (2) subjective comments elicited through debriefing questionnaires.

TEST SHIP DESIGNATION AND EQUIPMENT INSTALLATION

Once the implementation and test plan had been devised, COMCRUDESANT was informed of the plan and, through the fleet liaison function performed by DESDEVGRU, USS TRIPPE (FF 1075) was designated for participation in the study. The Commanding Officer and staff of TRIPPE received briefings concerning the program objectives, planned innovations, and data collection activities that would take place.

The FM equipment and materials were placed aboard ship. The carpetting, walk-off mats, pressure washer pumps, and trash compactor were installed by contractors.

Following a program orientation briefing, the skill/knowledge and attitude and motivation tests were administered to members of the Deck Division. The eight enlisted men (either nonrated seamen (SN) or Seaman Apprentices (SA)) selected and assigned to the FM team received initial training in the use and maintenance of the new equipment and materials. Responsibilities of the team and the new concepts of FM management, training, and operation were discussed with the team and the team supervisors. Data collection responsibilities (for man-hour recording, space inspections and training attendance) were delineated. It should

be noted that, due to replacements, sickness, or personnel transfer, a total of 12 men served as FM team members during various phases of the study. Only six served as team members during the entire study period.

DEPLOYMENT AND DATA COLLECTION

The test ship deployed for 6 months.

Task time data were collected daily for the deployment period. Completed JICs were turned in to the work center supervisor (leading Boatswain's Mate Chief) who retained them for pickup by data collection personnel.

Space inspections were made periodically by officers and the work center supervisor. FM inspection forms were completed and returned to DESDEVGRU personnel. Approximately three such forms were collected during the period of deployment. Training records were maintained by the work center supervisor, who recorded the dates each team member attended training sessions.

The test ship was boarded by a data collection team approximately midway through the deployment period. Skill/knowledge and attitude/motivation tests were readministered at that time. Several interviews regarding the progress of the study and effects of the innovations were also conducted.

Towards the end of the deployment period (after the test ship had returned to port), final administration of skill/knowledge and attitude/motivation tests was performed. Additionally, debriefing interviews and questionnaires were administered.

Data for comparison (control) purposes were collected on other FF 1052 class ships, including but not limited to USS BLAKELY (FF 1072), USS BROWN (FF 1089), USS HEWES (FF 1078), USS BOWEN (FF 1079), and USS PHARRIS (FF 1094). These data comprised estimates of task times and judgments of cleanliness and appearance of shipboard spaces. However, since the raters evaluating the control ships for comparison purposes with the test ship were not the same as those who had submitted the overwhelming majority of ratings aboard the test ship, no direct quantitative comparison was judged feasible in the analysis or interpretation of results.

A similar problem occurred with respect to task time aboard the control ships. Completed JICs for these were not available since the ships were not using the information management system. Instead, observers interviewed shipboard FM personnel to determine the amount of time spent on certain task aggregates. Thus, interview results could not serve as direct comparisons of FM task times in the analysis. These latter data are nevertheless considered useful and will be discussed later.

The raw data for the entire study was examined and analyzed. The next section presents the results.

1. FM man-hours were reduced from 20 to 40% due to FM innovations.
2. Spaces maintained by the FM team were generally rated as satisfactory or better with respect to overall appearance and cleanliness.
3. FM skill/knowledge of FM members increased.
4. Job attitude and motivation levels of FM members did not increase.
5. The overall FM program and various aspects of it generally received favorable ratings.

Comments: The following points should be noted with regard to this study.

1. The investigator addressed himself to major subsystems of the total FM system and, in fact, created certain subsystems. He did not confine himself to personnel performance (the output) alone. This illustrates that in performing system research it is necessary for the researcher to deal with substantially more of the system than individuals alone.
2. It was possible to carry out a controlled experiment within the operational environment, although certainly the amount of preparatory work was extensive--much more so than in almost all laboratory experiments.
3. The length of time required to gather data was extensive, far longer than for almost all laboratory experiments. In part this was because the operating system--the ship--has a prolonged mission cycle; in part because evaluations of major subsystems require substantially more time in order to get representative samples.
4. Because of the size of the system involved--an entire ship--it was necessary for the researcher to make use of test participants as data collectors.
5. Although sufficient control could be exercised over the major participants in the test, complete control was not possible.

b. Women Content in Units Force Development Test (MAX WAC), U. S. Army
Research Institute, Alexandria, Virginia, 3 October 1977.

PURPOSE: The purpose of this research was to assess the effects of varying the percentages of female soldiers assigned to representative types of category II and III TOE Units on the capability of a unit to perform its TOE Mission under field conditions. The objective was to provide empirical data to test the null hypothesis that specified increases in the proportion of women in selected units would not impair unit performance.

APPROACH: The basic concept was to test a total of 40 combat support and combat service support companies. These companies were broken down into eight companies each from five different types of units (Medical, Maintenance, Military Police, Transportation and Signal). Within each unit type the eight companies were designated as experimental, control, or calibration. Two experimental companies were to be tested twice, at varying fills of enlisted women (EW). The time between tests was to be six months. The control company was also to be tested twice with the EW fill stabilized for both tests. Five calibration companies were to be tested only once, with whatever percentage of women they contained. These companies established the range of scores one might expect, and some provided an opportunity for evaluators to gain experience before testing the experimental companies. The major statistical comparisons, however, were made between companies which were tested twice. The test design for the eight companies of each type unit appears as follows:

FILL LEVEL OF ENLISTED WOMEN FOR EACH TYPE OF UNIT

Test Season	Experimental		Control	Calibration	
	1 Co	1 Co	1 Co	2 Co's	3 Co's
Fall 1976	0%	15%	% as found	% as found	
Spring 1977	15%	35%	Same	% as found	

RESEARCH DESCRIPTION

a. Test Design.

Formulation of a scientifically sound research design, given the parameters imposed by "real-world" conditions, resulted in a methodology of somewhat limited scope but responsive to the basic question posed. ARI attempted to isolate the effect, if any, of different percentages of enlisted female soldiers on the performance of combat support and combat service support companies during a short-term (3-day) field exercise. It should be emphasized that, in accordance with the charter given ARI, attention was directed primarily to unit, not individual performance. Women who participated in the test were required to be MOS qualified. Furthermore, it was required that they be

assigned throughout the company. To test the major hypothesis of the project, it was necessary to determine whether the company could accomplish the myriad tasks which collectively make up its stated mission.

The core of the experimental design was a repeated measures (longitudinal) approach in which a company would act as its own control. Thus, the companies assigned to the experimental group were tested first at one level of female enlisted fill and about six months later at a different level of fill. To assess the effect of testing the same unit twice, the control group was to be tested during the first cycle of tests, the personnel stabilized as much as possible, and then tested again during the second cycle of tests. The remaining companies were tested once, about half during the first cycle of tests and the other half during the second cycle. This last group, referred to as the calibration group, served at least three purposes. Since there was no time, given the milestones provided to ARI, to pilot test the instruments and procedures that were to be used, by scheduling some of these calibration companies first, experience could be gained before the testing of the experimental and control companies began. Secondly, the range of scores, if not especially narrow, would allow statistical calibration of the scores obtained by the other two groups. Thirdly, since the percentage of women in companies varied, cross-company comparisons could be made between percentage of women in a company and ARTEP scores.

b. Test Instruments.

(1) ARTEP (Selected Tasks).

To assess company performance in the field, ARI was directed to use a standard operational Army test. The ARTEP is based on an analysis of the unit's mission and lists the various tasks the company must perform in accomplishing that mission. Guidance is provided for constructing a 3-4 day field exercise scenario to assess the company's ability to perform its mission.

The goal was to extract from each ARTEP a sufficient number of tasks to keep the company active as well as to require them to demonstrate competence in accomplishing tasks deemed especially critical to the unit's mission. The scenario had to weave these critical tasks, along with others, into a 72-hour exercise that would constitute a realistic test of all sections of the company with a minimum of task simulation. It was, of course, accepted that the threat imposed by an enemy--ambushes, aggressor attacks on unit perimeter, casualties to be processed by medical companies, etc.--required simulation.

A two-part scoring procedure was developed to provide more detailed assessments of company performance. Tasks and the sub-tasks were first rated on four separate factors. Table 1 lists these four factors and the definitions provided to the evaluators.

Table 1
Performance Evaluation Factors

Factor	Symbol	Definitions
Teamwork	Tw	Effective cooperation and coordination of effort between individuals working on a common task. (If test module or sub-task is performed by a single individual, teamwork is <u>not</u> assessed.)
Need for Supervision	NS	Each individual demonstrated appropriate skills, knowledge and abilities for task and requires only minimal level of supervision. Each individual carries full share of workload and demonstrates capability of working independently.
Timeliness	Tl	Task or mission accomplishment within a suitable or allowable length of time.
Quality of Work	QW	Mission accomplishment is judged with respect to the accuracy, correctness and efficiency of action and the quality of the product. How well was the job done?

In rating tasks and sub-tasks, the evaluators were instructed to use a three-level rating scale as shown below:

<u>Score</u>	<u>Basis of Rating</u>
1	Unsatisfactory
2	Satisfactory - Average to slightly above average
3	Outstanding

(2) Collateral Research Measures.

ARI did not have an opportunity, within the time frame specified for conducting the test, to pilot test instruments and procedures. As an aid to interpreting the test results, a set of questionnaires was developed to collect additional information, attitudes and opinions from the participants. These questionnaires were designed to provide insights into

organizational and individual factors that impact on the effect that content of women has on morale and performance in these combat support and combat service support units.

c. **Training Package.** A major concern, for the companies undergoing repeated testing with same scenario was the effect feedback from the first administration might have on the second test. It was felt that poor performance on tasks during the first test could cause the conscientious company commander to concentrate training time and resources to correct the deficiency before the second test. Two measures were taken to attempt to counter this possibility. In the first place, the design plan called for all twice-tested units to be given a 60-day training period prior to each ARTEP. The required female level of fill was to be attained before the start of the 60-day period. A training package was delivered to the company before the beginning of the training period; the package contained a detailed Letter of Instruction (LOI), the school-produced ARTEP and the summary of the scenario to be used on the field exercise. Additionally, arrangements were made for all reference material listed in the ARTEP (FMs, TMs, TCs, etc.) to be delivered to the company by pin-point distribution.

The five companies of each type tested once (calibration group) were given the same amount of time to prepare for the ARTEP and the same materials and information (Training Package). They were also required to maintain a training log in order to create comparable test conditions for all companies.

d. **Test Directorate.** (Comment: Note the necessity for having qualified specialists as evaluators.)

A Test Directorate was established, with a Test Director (COL) and a Deputy Test Director (LTC), consisting of five Evaluator teams (called Umpires in the OTP). Each team was to be headed by a branch qualified Team Chief, in all cases but one a Major, with command experience in that branch. The remainder of the team consisted of one branch qualified CPT, one combat arms CPT and one female CPT, branch immaterial. An administrative NCO (E8) and several civilian clerk typists completed the Directorate personnel. During the Fall test cycle, they were stationed TDY at ARI headquarters in the Washington, D.C. area. After the first of the year, about half of them returned to their home stations, while the other half remained in Washington. Those who had returned to their home station went TDY to each ARTEP location and periodically to ARI for conferences and to deliver completed instruments. Conduct of each ARTEP was under the direction of a local post evaluation team who were required to use the ARI-developed scenario.

Scenarios were written for the five types of companies to highlight the work of these soldiers. The scenarios

were written with three major considerations in mind. (1) Each was written in accordance with a SCORES mid-intensity European scenario. (2) Each was written to reduce the decision-making role of the company leadership. This was done to try to standardize the test procedures across all eleven ARTEPs. (3) Each scenario had to contain many tasks in addition to the critical tasks rated by the evaluators, in order to ensure that the whole company was kept occupied during the entire 72 hours. Although soldiers were not stressed or taxed to the limit, a realistic test required that there be little nonproductive time. In line with this philosophy, only genuinely malfunctioning equipment was to be repaired or actual messages transmitted. Simulation was used only when it was impractical to have the real thing.

b. Assignment of Women.

The Outline Test Plan defines the conditions governing the assignment of women in those units in which the level of fill was controlled. The most important consideration was that females be assigned in a large number of MOSs contained in each company's TOE; otherwise, the entire purpose of the test would be invalidated.

A second requirement specified in the OTP was that "all personnel available for duty at the time of the ARTEP shall participate in a manner appropriate to his/her MOS." The OTP directed that commanders not allow their companies to leave women behind in the company area during the ARTEP, "to handle essential administrative or urgent installation support--except for such reasons as illness or physical injuries." To ensure that the companies "don't leave the women behind," they were required to supply unit rosters and to account for all company personnel.

c. Control of Variables.

A field experiment of this magnitude involves so many variables which might impinge on the dependent measures (i.e., unit of performance) that control of all variables is extremely difficult, if not impossible. (Comment: Note the following important point which has been emphasized previously.) In the absence of direct control and of pilot work, one recourse is to measure (or record) as many aspects as possible of the conditions under which the tests are conducted and attempt to effect statistical control of these variables.

RESULTS

Type of Comparison		Change in Mean Score	t Statistic
Control (1st)	Control (2d)	Slight Decrement	Non-Significant
0%	15%	Slight Decrement	Non-Significant
15%	35%	Slight Improvement	Non-Significant
0%	35%	Slight Decrement	Non-Significant

Conclusion: Performance differences between the first and second ARTEP administration are small enough to be attributable to chance; an effect due to the change in content of women is not established.

Inference: Women soldiers up to percent tested do not impair unit performance during intensive 72 hour field exercises.

5. Laboratory Research

Feallock, J. B. and Briggs, G. E. A Multi-man-machine System Simulation Facility and Related Research on Information-processing and Decision-making Tasks (Report AMRL-TDR-63-48). Behavioral Sciences Laboratory, Wright-Patterson Air Force Base, Ohio, June 1963.

Comment: The following is an outstanding example of the development of a simulated system to be operated under laboratory control. This is an extremely complex system; it is possible to achieve similar results with simpler systems.

a. System characteristics:

(1) Team organization: A team of subject operators (Ss) consists of five members who bear the following titles: commanding officer, operations officer, terrain officer, structures officer, and vehicles officer; the latter three officers are all reconnaissance officers.

(2) Operational context: Subject operators are told that they constitute a reconnaissance-intelligence team whose mission is to establish the locations of critical installations in enemy territory by directing reconnaissance flights that will provide them with reconnaissance information, and by evaluating such information so as to determine actual and probable installation sites. Subject operators can initiate and control reconnaissance flights by issuing flight plans for as many aircraft as are made available by the experimenter; these aircraft may be used repeatedly. Reconnaissance flights over enemy territory are simulated in real time by a digital computer in accordance with subject-prepared flight plans. The simulation of a flight by the computer involves the following operations: (a) changing at 2-minute intervals the geographic location of the aircraft within the reconnaissance environment (see paragraph 4, below) in accordance with the course and speed specified on the flight plan; (b) establishing which environmental characteristics are "detected" during the flight by proceeding through a Monte Carlo routine for each 25x25-mile square cell of territory that the aircraft passed through for as many times as there are items of reconnaissance information in the over-flown

cells (typically four items per cell); (c) computing the amount of fuel used for each 2 minutes of flight, taking speed and altitude into account, and subtracting this amount from the previous residual (also "destroying" aircraft with insufficient fuel residual); (d) printing out results of the above operations every 6 minutes in the form of a reconnaissance report which constitutes the feedback input to subject operators. For each S who is assigned aircraft, feedback is limited to information gathered by only those aircraft for which he is responsible.

(3) Housing and equipment: Four rooms were devoted to this experiment. Work stations for two operationally independent teams of Ss were located in the "subject room." The teams were separated from each other by a heavy velvet curtain which served both to isolate the teams visually and to attenuate the sound level of conversation. The work station of each S consisted of table space, a chair, an intercommunication switchboard, and a two-earmuff headset with a boom-type microphone. The Ss were required to wear the headsets at all times during trials and also to conduct all conversations via the intercommunication system. The station connections of the intercommunication system were wired in a patchboard fashion so that the experimenter (E) could disconnect undesired communication channels. This system allowed Ss to have either private or conference conversations. The capacity of the system was 15 stations. Three of these stations were used by control personnel--two were used by operators called "communications officers" who read feedback information from computer reports over the intercom to subject operators, and one was used by E for monitoring conversations and issuing instructions. All voice communications of Ss were recorded on tape directly from the communication network.

(4) The reconnaissance environment: The source of data for Ss was "enemy territory" as viewed by airborne sensors. The sensor perspective of enemy territory was simulated with an abstract map on which identifiable characteristics of enemy territory were entered by E in unit areas (25x25 miles) of a total area (750x375 miles) comprised of 510 unit areas (30x17). Identifiable characteristics of a unit area or cell were its terrain (plains, forests, or hills) and the structures (warehouses, storage tanks, or antennas) and vehicles (general-purpose, tankers, or weapons carriers) located in it The map of enemy territory just described was never seen by Ss, but characteristics of individual cells became known to them under appropriate conditions of air reconnaissance.

(5) Information-processing or intelligence rules: Each subject operator was given the rules according to which installations were assigned to cells to guide him in directing reconnaissance flights. The alternative to giving S the rules

was to allow him to infer them from repeated observations and analyses as is done in real life. Such an approach would not only be prohibitive in cost but would magnify individual differences and thus reduce the power of the experiment.

The intelligence rules were made available to Ss in the form of tables which, for each combination of 1, 2, 3, and 4 characteristics known for a cell, reflected the probability that an installation was located in the cell. The probability for chance occurrence was .15; the other probabilities ranged from .00 to .67. In general, as more of a cell's characteristics became known, the better the evaluation of the cell became; i.e., the S could be more sure that this was a cell that should be reconnoitered further or a cell that was to be avoided.

(6) Operator tasks: Two subject teams of six members each had identical organizations and operated simultaneously but independently. The responsibilities of the various officers were as follows:

Commanding officer--to maximize the number of installations detected within each duty period (a 4-hour experimental trial) by the following means:

(a) by evaluating and guiding the performance of subordinates, namely all other members of the team;

(b) by specifying flight plans for installation aircraft to the XO that take maximum advantage of the reconnaissance evaluations or recommendations forwarded by ROs and that utilize a minimum of fuel;

(c) by requesting of ROs evaluations of specific cells when doing so could benefit the planning of flights.

Executive officer--to be of maximum assistance to the CO in implementing the preparation of suitable and efficient flight plans:

(a) by drafting flight plans for the CO;

(b) by computing fuel requirements for flight plans of the CO;

(c) by submitting approved flight plans.

Reconnaissance officers--to identify cells and clusters of cells for the CO which have especially low and especially high probabilities that installations are located in them:

(a) by specifying, drafting, and submitting flight plans for reconnaissance flights to the OpnsO;

(b) by coordinating with the other ROs in the planning of flights;

(c) by recording and processing information on cell characteristics to determine the probability that an installation is in each cell;

(d) by exchanging information on cell characteristics with the other ROs;

(e) by selecting and reporting cells of high and low probability value to the CO.

Operations officer--to be of assistance to the ROs in implementing the preparation of suitable and efficient flight plans:

(a) by computing fuel requirements for flight plans of the ROs;

(b) by forwarding satisfactory flight plans to the XO.

(7) Response options and constraints: The CO was assigned six aircraft that were equipped for detecting and identifying installations but not for establishing the characteristics of cells; therefore, above-chance success in detecting installations with these aircraft depended upon the identification of highly probable installation locations from information gathered by other aircraft. The CO's aircraft were to be used for "inspection," then, rather than for reconnaissance. The ROs were assigned nine aircraft in all for establishing the characteristics of cells but not for detecting or identifying installations. The distribution of the nine reconnaissance aircraft among the ROs was determined largely by the design of the experiment and, in fact, constituted a major independent variable. The aircraft differed in the sensors they carried and, therefore, in the cell characteristics that were most readily sensed during their use: although all aircraft could be used to detect all characteristics, each was well equipped for the detection of ground features and for one of the other three sets of characteristics. For this reason the reconnaissance aircraft were distinguished by the names "terrain sensor," "structures sensor," and "vehicles sensor;" there were three aircraft of each type. Each RO was provided with at least one aircraft of a type conforming with his specialty, but never were all of his aircraft of that type.

When aircraft were assigned to ROs, five were assigned to one air base and four to another. Flights were required to originate and terminate at the aircraft's base of assignment.

The fuel capacity of each aircraft was fixed at 40,000 pounds. The rate at which fuel was used in flight was made a function of flight speed and altitude. Choices of three speeds (375, 562.5, and 750 mph) and of six altitudes (1, 2, 5, 10, 20, and 40 thousands of feet) were available. The Ss were given both tables and graphed functions that reflected fuel consumption rates for different combinations of speed and altitude in terms of both pounds per minute and pounds per mile. To effect

a reconnaissance flight S submitted instructions on a standard form (a flight plan) to the computer center, specifying a particular aircraft to perform the flight, the course of flight, and the speeds and altitudes to be flown. As many as ten check points (geographic coordinates that defined straight-line flight legs) could be specified in a flight plan, and fuel requirements had to be computed for the distance traveled between each successive pair of check points. . . . The penalty for fuel exhaustion in flight was the loss of the aircraft. This penalty was not often levied as Ss were given periodic information on fuel expenditure for flights in progress and Ss could change flights while they were in progress.

The importance of speed and altitude factors was that both determined in part the probability that cell characteristics and installations would be detected. Detection probability varied inversely with both altitude and speed. The exact mathematical relationship between these variables was not made known to Ss. Since problem duration was a task constraint, flight speed was an especially important factor because it determined the number of cells that could be reconnoitered per unit time. Speeds were established by E which would allow Ss to reconnoiter cells at the rate of 1, 1-1/2, or 2 cells per 2-minute interval.

b. Method and procedure:

(1) Operational procedure and experimental schedule:
Prior to the conduct of each experimental trial a map of the reconnaissance environment was selected from a map library in accordance with the experimental design, encoded on punched tape, and entered into storage in the digital computer. The experimental trial was begun typically at 1:00 p.m. on a week day and continued until 4:45 p.m. A trial was divided into two phases: a planning phase of one-half hour and an operational phase of 3 hours; the operational phase was interrupted at midpoint for a 15-minute rest break. The Ss were seated at their respective work stations during both phases of a trial.

Six flight plans could be submitted during the planning phase for flights to commence at the start of the operational phase. . . . When the operational phase began, flight simulation was initiated on the computer and was carried out at twice real-time rate. At the end of each evaluation cycle duplicate printed copies of information prepared by the computer were available; one copy was retained for experimental records and the other was used in the following way. The flight progress information was separated from the reconnaissance data and delivered to the XO and the OpnsO in hard-copy form. The reconnaissance data were read to Ss individually by the communications officer via the intercommunication network. Each operator was told only those cell characteristics that were detected by the aircraft assigned to him and each such report required an average of about 2 minutes. The installations detected on flights of the CO were reported

to him only. The operators were instructed to record all reconnaissance information on maps as it was reported to them. That portion of each evaluation cycle which was not used by the RO in receiving the reconnaissance report was devoted to evaluating the data, exchanging information verbally on cell characteristics with the other ROs, coordinating the planning of future flights with other officers, preparing flight plans, and reporting cells with exceptionally low and high probability values to the CO.

Flights conducted for the detection and identification of installations were carried out to completion even after the trial terminated, and the subject team was credited with all installation detections made during or after the trial. Final knowledge of results (total number of installations detected) was provided to the team just prior to the beginning of the next trial.

All Ss who participated in the experiment were given 2 weeks (10 trials) practice in their assigned experimental positions prior to the experiment under the control condition of equal distribution of load. Practice trials were conducted in the same manner as experimental trials, except that they were marked by more frequent inquiries on the part of Ss regarding experimental limits and legitimate procedures. During both the training period and the experimental period E was available to answer questions via the intercommunication network.

(2) Subjects: Eighteen male students who were enrolled at Ohio State University comprised two subject teams, each of which consisted of a permanent contingent of three members and two alternating crews of three members. None of the Ss had experience with military operations of the sort being simulated, and only three Ss had served as system operators in previous experiments.

(3) Experimental design: The design of the experiment provided for an analysis of performance measures in terms of the following factors (levels in parentheses): load distribution (4), trials (5), teams (2), and crews within teams (2). For load distribution a fixed number of reconnaissance aircraft (9 aircraft) were distributed among the three ROs equally (3-3-3) under the control condition and unequally (1-3-5, 1-1-7, 1-4-4) under the experimental conditions. The second factor was defined by five experimental trials performed in a block by each crew of each team under each condition of load distribution. The third factor was defined by two teams each of which operated with six members at a time. The fourth factor was defined by the two crews of ROs for each team. The three types of ROs (terrain, structures, vehicles officers) were nested in teams and crews. The order in which teams and their crews performed under the different conditions of load distribution was established by a Latin-square design which is shown in table 1.

The two teams of Ss were run simultaneously on identical, experimentally independent facilities.

Table 1

Order of Performance Under Experimental Conditions (IPAC V)

Team	Crew	Levels of Load Distribution			
		3-3-3*	1-3-5	1-1-7	1-4-4
1	A	1	2	3	4
	B	2	4	1	3
2	A	3	1	4	2
	B	4	3	2	1

* Numbers of aircraft assigned to different ROs of a crew.

(4) Dependent variables: Total system performance measures: The measures taken on the total system's performance fall into two categories--output measures and cost measures. Efficiency indices were computed by combining these measures. The number of installations detected per trial was the only output measure taken. The cost measures taken were (a) pounds of fuel used in reconnaissance, (b) number of reconnaissance flights flown, and (c) reconnaissance aircraft use time in minutes. Performance efficiency was computed three ways by dividing the output measure by each of the cost measures.

Subsystem performance measures: The two dependent variables chosen for investigating adjustment at the subsystem level were (a) the number of cells for which information was communicated to other ROs by each RO (cells reported), and (b) the number of cells for which information was received from other ROs by each RO (cells received). To ascertain which RO or ROs were performing most of the final evaluations of cells, a third dependent variable was identified which was the number of high-value cells reported to the CO by each RO (cells recommended).

6. Normative Data Gathering

It is the author's feeling that the greatest present need in PSM research is the gathering of normative data describing personnel operations during system functioning. These data can then be used for purposes of predicting the performance of other personnel in other (similar) systems in which the same or comparable tasks are to be performed.

Unfortunately, laboratory research and even PSM studies making use of CE designs cannot produce normative data for several reasons: because they

concentrate only on a sample of or the more extreme values of the variables being tested; because they exclude significant interactive variables; because they test only special situations of interest to the hypothesis-maker. In effect, CE researchers often ignore the operational environment.

The lack of interest in normative data is all the more surprising because gathering such data is on the whole much simpler and much more rewarding than are CE studies.

Two methods can be used for gathering normative data: manual and automatic. The first method has a long history, but has been supplanted (sometimes) by the automatic method in which recording equipment is attached to the man-machine interface. The manual method which employs observers and system personnel as reporters necessarily infuses a subjective element into the data secured, but can cover a much wider range of system tasks than can automatic recording. The automatic method which largely eliminates the subjective element is restricted to those personnel responses which can be tapped by the recording equipment. Consequently automatic data gathering systems are useful primarily for highly molecular psychomotor outputs (e.g., button-pushing, as in the OPREDS system developed by the Navy Ocean Systems Center to measure control responses in the Navy Tactical Data System). Both methods are useful, depending on the data being gathered.

The kind of normative data the author would wish to have (or at least the output of those data, once evaluated) is represented by the following example which is wholly imaginary, since no such compilation of data exists anywhere:

<u>Task Activity</u>	<u>Probability of Correct Performance</u>
Turn rotary selector switch and observe.	0.9950 ¹
CRT signal quality.	0.9972 ²
	0.9988 ³
	0.9965 ⁴
	0.9960 ⁵

Notes: (1) Apprentice level operator; no negative task/environmental conditions.

(2) Moderate skill level operator; same conditions as in (1).

(3) High skill level operator; same conditions as in (1).

(4) Moderate skill level operator; pressed for time.

(5) Moderate skill level operator; multiple task requirements.

A series of tables would describe all major tasks required by at least the more common systems, under major qualifying conditions such as personnel skill level, competing task requirements, environmental stresses, etc. Such data have been described in terms of "human performance reliability" (Meister, 1964).

a. Manual Method

Grings, W. W. et al. Shipboard Observations of Electronics Personnel: Detailed Descriptions of Observational Techniques (Technical Report 2). Department of Psychology, University of Southern California, January 1953.

The purpose of this study (one in a series) was "to obtain a description of the electronic situation as it exists in ships of the destroyer class." A number of methods were employed.

(1) The job questionnaire. Respondents described themselves and various aspects of their jobs. Forty-four items covered such things as schooling, experience, duties and materials (tools, test equipment, etc.). The questionnaire was completed by ship personnel on their own time. Most frequent form of analysis was the simple frequency count.

(2) The diary. The observer recorded in a time sequence everything that took place in a predetermined area of observation. Areas of observation were defined by the particular orientation (the man, the trouble, the place) the observer assumed. Man-oriented observations required the observer to note everything a man did in a given period of time. Trouble-oriented observations consisted of observing everything that took place in the process of returning malfunctioning equipment to satisfactory status. In the place-oriented approach, the observer recorded the activities of electronics personnel in a particular ship space. Observations were recorded with paper and pencil or by tape recording. The observer endeavored to remain in the background, although

obviously ship personnel were aware of his presence and why he was there.

(Comment: Manifestly what one looks for depends on the observer's concept of the important variables in the performance.)

(3) Ability Requirements Scale. A rating scale was prepared to determine the qualifications and abilities needed for a given electronics job. The scale contained a number of different traits which supervisors were asked to rate in terms of importance for the job. Nineteen traits were rated individually on a 5-point Likert-type scale.

(4) A general questionnaire sought opinions on important problems in electronics.

(5) Interviews (10-90 minutes long) were open-ended and semi-structured.

(6) Critical incidents were sought to determine the behavioral components in a situation which added to or subtracted from its effectiveness.

(7) A record summary form was completed with information contained in standard repair records kept by technicians aboard ship. The purpose of these data was to obtain a large volume of data on the kinds of repairs made: equipment repaired, nature and cause of the trouble, and kind of repair made.

(8) A training questionnaire of 211 items designed to determine the usefulness of various curriculum topics to the technician's job.

(9) A repair record completed by technicians to describe the repair work they performed and how they went about it.

(10) The log was the observer's narrative account of his observational trip (the observer's activities, time and date of occurrence, etc.). This was useful to describe activities not otherwise recorded by the other methods. It also contained comments on the value of the methods employed as well as a description of the general maintenance situation aboard ship.

(11) The checklist presented supervisors and technicians with a list of activities which various rates and ratings might perform on the job. Respondents checked the rates and ratings of men whom they actually observed performing a particular activity on a specific equipment.

(12) The card sort method required the sorting of a deck of job statements into a number of predetermined activities. The sorter indicated which activities he had performed as part of his job, how important and how often, where he learned these activities, and the amount of electronics comprehension and skill required for these activities.

Note the variety of the methods that can be applied and that these methods are most useful with more motor, perceptual and cognitive tasks, such as electronics maintenance. Because these methods are so dependent on observer competence, they require extensive training of the observer and scheduling of his measurement activities. The amount of data collected depends on the opportunities to use these methods. Those methods which rely on the observation of particular occurrences will yield no information if the activity to be observed never occurs. Because the observer interacts very directly with system personnel, it is necessary to inform them regarding the purposes of the research, possible effects upon the crew, etc. The problems involved in the analysis of a large amount of descriptive data are many; the information in its raw form has little value. The analyst must extract from its mass the main points or principles it depicts.

Other normative data gathering systems depend on automatic sensing and computer processing. The example described below illustrates such a system.

b. Automatic Method

Eatock, B. C. A Portable Interactive Data Acquisition and Analysis System for Driver Behavior Research (DCIEM Technical Report 77X30). Defense and Civil Institute of Environmental Medicine, Toronto, Canada, 15 June 1977.

The Road Safety Unit of Transport Canada has a general requirement to perform measurements of driver behaviour under actual driving conditions. DCIEM was tasked to provide a car-portable system to provide these measurements.

The general requirements for the instrumentation system were:

1. True portability. It was to fit most North American cars, mid-size and up. Installation time of less than 24 hours was desired.
2. Low power consumption.
3. Modular design.
4. On-line data analysis capability.
5. On-line control over experimental procedures.

The performance of the system is outlined in Table 1. The system also has the capability of transforming basic measures and/or combining them on line, to form complicated derivative measures.

The central component is a DEC LSI-11 microcomputer, a compact machine which consumes relatively little power. When interfaced with appropriate peripheral devices and operated with commercially available software, it behaves like a conventional minicomputer. In the present configuration there are 24K (24,576) words of metal oxide semiconductor (MOS) memory, with the option to add another 4K module.

Random access bulk storage is provided by an RX01 floppy disk unit. The storage medium is a preformatted flexible diskette, or floppy disk, which can store 256K eight-bit bytes of information. The 'floppies' are used for storing data collected during the course of an experiment and providing permanent records of programs.

The other peripherals which provide communication between the operator and the computer include: a keyboard terminal with a single line gas-discharge display, a thermal printer for providing limited hard copies of numerical data, and two digital clocks. One clock reads the time of day in hours, minutes and seconds, while the other provides a millisecond readout. The two clocks can be used to time events under program control. A hardware bootstrap facilitates system initialization when the computer is turned on.

The LSI-11 communicates with the measurement transducers by means of a data acquisition module which samples up to 24 differential analog channels and performs a 12 bit analog to digital conversion for each. Under program control the computer can provide two channels of analog output, using two, 12-bit, digital-to-analog converters. In addition, there is a provision to accept either discrete inputs from digital transducers or to generate discrete signals for controlling external devices (e.g., light signals).

Table 1

Measurement Parameters

Parameters	Range	Accuracy ¹	Sampling Rate (Hz) ²
Time (Relative)	practically unlimited	0.001 s	100
Speed	0-120 mph.	3%	10
Distance ³	practically unlimited		
Acceleration X	0 to $\pm 2.5g$	0.5%	20
Y	0 to $\pm 2.5g$	0.5%	20
Z	0 to $\pm 2.5g$	0.5%	50
Steering	± 20 (fine)	0.5%	50
Wheel Position	± 360 (coarse)	0.5%	50
Accelerator	0 to full	$< 1\%$ ⁴	50
Brake Pedal Force	0 to 300 lb.	1%	50
Lateral Position	$\pm 2m$	0.5%	50
Discrete Driver Responses			interrupt
Analog Responses		1%	10

- Notes: 1. accuracy given in % of full scale, except time
2. typical
3. distance is integrated from speed
4. nonlinear measurement, accuracy, determined by calibration table

Speed is measured by a microwave radar Doppler speed sensor and a measure of distance is derived by summation. Coarse and fine readouts of steering wheel position are obtained by measuring the rotation of two potentiometers. Accelerator position is obtained by measuring the displacement of the linkage at the carburettor with a linear displacement transducer. Acceleration along three axes is measured by three, sensitive, low-frequency, force-balance accelerometers, mounted orthogonally. Brake pedal force is determined using a commercially available force transducer. A sophisticated optoelectronic lane tracker, specially designed by Human Factors Research Corp., measures the lateral distance of the vehicle from the roadway centreline.

The system is powered by a high current (130 amp) alternator which replaces the vehicle's original alternator. A reserve battery is included. A nominal twelve volts is supplied to a 12V/12V converter, 12V/5V converter and a 115V/60 Hz inverter to provide well regulated power. Maximum power consumption in the present configuration is less than 800W.

The total weight of the system (140 kg) is evenly distributed throughout the vehicle. No single component weighs more than 25 kg.

THE INFLUENCE OF GOVERNMENT ON
HUMAN FACTORS RESEARCH AND DEVELOPMENT

David Meister

US Navy Personnel Research and Development
Center, San Diego, California

How much Human Factors (HF) research and development (R&D) is performed in the United States depends in large part on how its infrastructure functions. That infrastructure consists of relationships among the executive and legislative agencies of government, their R&D laboratories, R&D contractors and HF practitioners in industry. (The governmental agencies referred to are: the military services of the Dept. of Defense, Army, Navy and Air Force, and their staffs; civilian executives like the Assistant Secretary of the Navy for Manpower, Personnel and Logistics Affairs; and Congressmen and their staff assistants. Although other agencies like the Dept. of Transportation engage in HF R&D, the amount they support is comparatively minor. The R&D laboratories are those like the author's laboratory that are part of the individual military services).

The purpose of this paper is to explore the relationships within the infrastructure as perceived by the three major participants: contractors; laboratory managers; and practitioners. (No attempt was made to secure information from the executive and legislative agencies because the probability of their responding seemed quite low.)

There are two major reasons for exploring this behavioral infrastructure: (1) Knowing more about the relationships involved may in the long run help to direct behavioral R&D into more effective channels; (2) The exploration offers insights into what might be termed the sociology of research, in particular how originally "pure" scientific intentions can be modified by the economic and political context in which these intentions must be realized.

Data to explore this infrastructure are not easy to secure. One deals here with practices and motivations which are usually shrouded from public view. It is possible however, to make a start by asking the people involved what their opinions are -- in short, to develop and administer

questionnaires addressing these topics to the three participants.

The author developed the questionnaire items based on his experience as contractor, laboratory worker and practitioner. On the basis of his knowledge of those most senior, most experienced and most influential in each category, he selected those to whom the questionnaires were sent. The samples cannot therefore be considered a random sample and therefore representative of the total population of each category; but it is felt that the responses are from those most qualified to make judgments on the subjects addressed. Because the selection criteria were restrictive, the number receiving the questionnaires (one for each category) was limited: 34 research contractors, 38 laboratory managers, and 41 practitioners. Of these 26 contractors, 30 laboratory managers and 27 practitioners responded, for an overall percentage return of 73%.

All respondents were promised anonymity but all except a few signed their names to their answers. The roster of those replying includes the most prominent people in the HF discipline.

The questionnaires employed statements which the subjects had to check on a 5-point Likert-type scale describing attributes like frequency (always, usually, sometimes, rarely, never) or satisfaction (very satisfactory, satisfactory, minimally acceptable, somewhat inadequate, completely inadequate). For example, "The outputs of contract research are generally: very satisfactory, satisfactory, minimally acceptable, somewhat inadequate, completely inadequate." Items for practitioners used 4-part scales (strongly agree, moderately agree, moderately disagree, strongly disagree). For other types of information such as percentage of work performed for the government, quantitative estimates were requested.

One thing that must be emphasized is that the conclusions described in this paper are based on the perceptions of the respondents and may therefore be only partially valid. Perhaps because of the subjective element there is great response variability. Another reason for this variability is that the different types of HF organizations within each of the three categories perform their functions in somewhat different

ways. For example, in industry the percentage of government-funded R&D performed by an HF group may vary from 0 to 100, depending on the role of that group within its company and what the company itself does. Similarly, in government laboratories, the percentage of R&D performed inhouse as against that contracted out also varies from 0 to 100%. Only among contractors do we find some consistency: The mean percentage of their R&D work funded by the government is 86%, only one contractor reporting as little as 50%. Where HF functions differ as they do within each respondent category, the answers one receives to individual questions will also vary. To clarify the answers respondents were asked to comment on their responses; the author will also comment on the conclusions described below to attempt to explain apparent inconsistencies.

Contractors

Government support is the necessary ingredient of all behavioral contractors and of most HF groups within industry. A few of the latter might be able to survive by working on purely commercial system development projects; but these would be a very small number. It is commonly held by practitioners that commercial industry (as opposed to governmental system developers) has little or no interest in HF; industry accepts HF only because the latter is subsidized or required by government on weapon system development.

Contractors receive comparatively little support (12% of the total) for basic research, if this research is defined as "studies of a theoretical or methodological nature, having wide generality and elicited by a general problem" (all quotes are from the questionnaires). Most of their support (61%) is given to applied research which "seeks to solve specific operational problems". 25% of the total is given for development defined as "the construction of an object or procedure, e.g., a manual, to be used by a specific operational system or type of system".

47% of contractor R&D is secured by bidding on competitive contracts. Surprisingly, however, 29% of funding is provided by unsolicited contract and 25% by sole source contracts - neither of which is competitive. What this means is that some Dept. of Defense agencies like the Office of Naval Research (ONR) or the Advanced Research Projects Agency (ARPA) disburse all their funds on a non-competitive basis, whereas other agencies like the Air Force's Human Resources Laboratory (AFHRL) disburse 90% of their funds on a competitive basis. Those government agencies disbursing funds non-competitively tend to emphasise basic research and to support university organizations more heavily; contractors must live with competitive contracts. The question one must ultimately ask is what the effect of this competition is on the quality of behavioral R&D results, and whether one should consider research in the same vein as buying nails.

55%* of the contractors responding considered that the research problems they attacked were either very important or highly significant. 42% felt that these problems were moderately important. The latter are apparently somewhat tepid about the importance of the research problems they address.

Contractors feel that they sometimes (64%) or usually (16%) have an opportunity to influence the selection of the research topic they are funded to pursue, and laboratory managers tend to agree that potential contractors sometimes (42%) or usually (10%) suggest a research topic that is later funded. This should be related to the 29% funding of unsolicited contracts which are presumably suggested by the contractor. However, this influence exists primarily for basic research in which the direction of the research is less structured. When contractors were asked whether "governmental agencies do, in fact, fund the most important research/development topics?(use own criteria)", 23% disagreed with the statement, 69% agreed moderately (suggesting a residue of scepticism) and only 7% strongly agreed. When the same question was asked in a different way, "do you agree that most governmentally funded behavioral research is directed to important topics?", 53% moderately agreed and 23% strongly agreed.

Behavioral problems of a system nature are inevitably interwoven with hardware and managerial elements. Consequently it is reasonable

* all values specified are percentages of the total number of subjects responding to a particular scale value on a particular question. Naturally not every respondent answered every question.

to ask contractors whether the problems that are funded to research can be solved solely by behavioral research. Most (47%) thought it could be sometimes, 30% usually or always; 21% rarely or never. The explanation for this dispersion of responses may be the nature of the research performed by individual contractors. If the problems they attack are mission-oriented, they can rarely be solved without involving hardware and management processes. If the problems are solely behavioral (as might well be the case with more basic research themes), the research can be effective on its own.

In contract research the one who pays for the research can dictate the way it is to be done, even though he does not perform the research himself. The question is whether, in performing government-funded R&D, the methodology is imposed on the contractor by the customer. 48% of the contractors indicated that this happens sometimes, 24% usually. However, 28% answered rarely or never. The explanation of these discrepant responses seems again to be the nature of the research. If the research is basic, the methodology is rarely imposed; for exploratory and advanced development projects the customer is much more likely to specify the methodology. There is also variation from service to service and even between laboratories in the same service.

A common contractor complaint is that behavioral R&D funding is usually inadequate. There is the usual response dispersion in connection with this complaint: 40% of the contractors think it is usually sufficient, 40% sometimes sufficient, 20% rarely so. Obviously funding is not always adequate. Among comments made concerning this item was one that government administrators are familiar with other R&D costs but not with that of behavioral R&D. Because of the indefinite nature of research the contractor tends to underestimate because he cannot see all the tasks that will arise. Because the competitive bidding process emphasizes low bids, contractors tend to underestimate in order to win contracts. After they win, they find that funding unrealistic. The saving grace is that one can always "tailor the job" to the money involved; one can do a bit less or in less detail.

Government contract manuals emphasize the importance of writing the Statement of Work (i.e., SOW, the specifications of the government's research requirements) as clearly and understandably as possible. Manifestly,

if there is confusion the contractor cannot give the government its money's worth. In this connection, only 40% of contractors felt that the SOW was usually clear; none believed it was always so; 60% felt that the SOW was only sometimes or rarely understandable. Following the same theme, only 44% felt that government R&D objectives were clear at the start of the research or development. 56% felt that such objectives were only sometimes or rarely clear. The lack of clarity makes it necessary for the contractor to spend precious time probing the contract monitor's mind. Since this probing will not always be successful, a certain percentage of R&D is destined to be unsuccessful. It was pointed out that unless one has had prior contact with the contract monitor, it is often difficult to understand the SOW fully. This situation varies from agency to agency. Because of legal regulations governing procurements, the government representative is inhibited, during the procurement process, from indicating fully the kind of effort desired.

Contract researchers are generally confident of the capabilities of their discipline. 32% felt that the objectives for which behavioral R&D is funded rarely exceed the technological capabilities of behavioral science. 64% felt that they did sometimes; only 4% said usually. How one answers depends again on the research task; there may be no problem performing a task analysis, but there are great difficulties in estimating the behavioral consequences of a nuclear attack.

If funding is a problem to contractors, so is the time frame within which behavioral R&D must be performed. The overwhelming majority (84%) of contractors feel that the time frame imposed by government on R&D contracts is too short to do an effective job. One of the problems is that time requirements may be related to fiscal funding periods rather than that actually needed. Because funding is tied to annual governmental budgets, many contracts must be organized around a 12 month or shorter period. The consequence is that the research job is cut to fit the milestones provided by the customer. It is, moreover, difficult for the contractor to predict the time required to do a job; often this can be done adequately only after getting into the project. However, the scope of most research projects is flexible and can be tailored to the time allotted (within limits of course). One contractor commented that the government's time estimates are pretty good provided that no problems develop and the contractor can devote 100% of his time to the project in question. However, these last two conditions rarely exist.

The political nature of some behavioral R&D (particularly evaluation research) is a matter of concern to some contractors. 45% felt that they are sometimes or usually expected to slant their conclusions to the implicit or explicit desires of the customer. On the other hand, 55% rarely or never encounter pressure. When pressure exists, it is subtle of course; contractors are never explicitly ordered to subvert their work but they know what the customer wants and the penalties for refusing to go along with him. Often it is a matter of wording a report so that a slightly different impression (e.g. more positive) is received; such pressure is much less common in basic research.

There is also a problem of implementation. If implementation is taken as a criterion of utility (a point of view taken by Congress and the General Accounting Office), much behavioral research has dubious credentials. 23% of contractors believe that their recommendations are rarely implemented by the customer and 58% believe that they are only sometimes implemented. Contractors in general do not feel that any inadequacy in their R&D is responsible for this lack of implementation. They view it as a lack of will, inertia, resistance to technological change or political obstacles.

Contractors also cast a somewhat jaundiced eye at the capability of their governmental monitors. 28% felt that their opposite numbers were rarely as competent in R&D as they should be. On the other hand, 44% felt that government monitors were sometimes competent; 28% usually or always. If their feelings mirror fact, the saving grace is that, as was pointed out by several respondents, monitors need not be competent because it is the contractor who must actually do the research and who must therefore be competent. But how is the monitor to know if the contractor is doing a good job unless he is as knowledgeable as the contractor about the research question at issue? This factor of governmental competency may well be more important in basic research than in mission-oriented R&D because in the latter the government monitor has the advantage of familiarity with the problem.

Considering the operational environment in which most contractors work and in which stringent scientific controls are difficult to impose, it is somewhat surprising that almost half the contractor sample considered that it was always (3%) or usually (42%) possible to apply academic standards of research rigor to these situations. On the other hand, half felt that it was only sometimes (23%) or rarely (30%) possible to do so. The answer seems to be in the kind of research situation permitted by the research

topic. As one contractor put it, "the critical factor is what the sponsor wants to achieve and the rigor with which he wants it". It may be possible to be rigorous if the problem can be fragmented, but this may not be possible.

Laboratory Managers

Laboratory managers indicated that on the average about 60% of their organization's R&D is contracted out, but there is wide variability,* some organizations contracting their total R&D, others none. The mirror image of this is that about 40% of governmental R&D is performed inhouse. Only 47% of contract research involves competitive bidding, which means that much R&D is not open to every contractor. 33% of the research contracted out is sole-source. The amount of basic research, exploratory and advanced development performed is respectively 17%, 27% and 41% (which accounts for 85% of laboratory activity; presumably the rest is paper shuffling). 43% of laboratory research directly supports system development.

As to whether government personnel prefer to contract out their R&D rather than perform it inhouse, 23% of the sample preferred contractor research; 40% preferred inhouse activity and 37% preferred a mix of both as the needs of the situation dictated. There is thus some preference for inhouse research, but not overwhelmingly so.

Although behavioral R&D is supposed to be heavily competitive because of governmental policy, over half the governmental sample (57%) report that their organizations tend to stay with the same set of contractors over a series of follow-on contracts. 25% say that they do so sometimes, 17% rarely. As one respondent indicated, the number of "good" contractors is limited. It may also be a function of how specialized a research topic is; if the topic is highly specialized, only a highly specialized contractor (and there are a few) is equipped to handle the business. Familiarity with certain contractors may lead to an unconscious preference for working with them.

Government personnel feel that they have great influence over research they fund (15%, very much; 61%, much; 23%, some). This is

* Standard deviations were calculated, but are not presented here because they merely indicate high variability which is already acknowledged.

logical but it raises a methodological and ethical question: should or can someone who is not performing the research exercise a determining influence over that research? Can one actually "buy" research?

Even further, do potential contractors ever suggest research topics which are later funded? If so, it would indicate that they are not completely powerless in directing the course of R&D, as well as being largely responsible for performing that R&D. 42% of the laboratory managers indicate that contractors rarely or never influence a research topic; an equal number say they do sometimes; and 13% say usually or always. The explanation is that when the research suggested is basic, the contractor has a much better chance that his suggestion will be accepted; if the R&D is exploratory or advanced development there is much less chance. The probability of an unsolicited proposal being accepted is generally low unless the contractor has intimate knowledge of the government's R&D plans and fits his suggestion into those plans. Sometimes, if a particular contractor's expertise is desired, he may be "fed" an idea which he refines into an unsolicited proposal, but this is rare.

Where then do research topics come from? There are four possibilities: staff of governmental agencies (e.g., The Dept. of Defense, The Military Services); laboratory managers; inhouse laboratory personnel; contractors/consultants. These categories describe a descending power structure, with agency staff having most influence and contractors least. Estimates from the managers bear this out: 53% indicated that research suggestions come primarily from agency personnel, 25%, laboratory managers; 21%, inhouse laboratory personnel; and 3%, contractors/consultants. The R&D areas pursued are determined therefore primarily by agency personnel, many of whom are not behavioral specialists, so one can legitimately ask whether they are qualified to provide R&D direction. Laboratory managers presumably have the expertise to provide direction but do no R&D themselves. The contractors who carry out much of the R&D have a limited amount of influence. One can hypothesize that laboratory managers and other behavioral specialists influence basic more than applied research, because the former requires more expertise than most non-specialists have. However, as the R&D topic becomes more mission-related, higher agency levels exercise a disproportionate impact.

Government personnel also quarrel with the funding provided for their R&D projects and if anything are slightly more negative about the situation than contractors. 23% feel that funding is either somewhat or completely inadequate; 43% feel that it is minimally adequate; only 33% feel that it is adequate. In addition to its being inadequate, funding is highly unstable. Decisions on which R&D program elements are to be funded at what levels are disproportionately influenced by external pressures and considerations of visibility.

Despite all this, the outputs of contract and inhouse research are generally considered satisfactory. Contract research outputs were considered very satisfactory by 6% of managers, satisfactory by 72% and minimally acceptable by 20%. Inhouse research follows the same general pattern, although 7% feel that these outputs are somewhat inadequate.

If one is concerned about the adequacy of governmental planning for R&D, exactly half the governmental sample indicated that an inhouse program plan was always required by laboratory management before a project is funded. However, there is great variability among laboratories; the other half of the sample was almost equally divided among usually (14%), sometimes (17%) and rarely (17%) required. One respondent indicated that these program plans are "managerial eye wash", only tokens that are not evaluated by scientific criteria. Detail in these plans is highly variable. Finally, some government personnel resent what they see as the "paper mill" behind the requirement.

By a vast majority (88%) governmental personnel feel that the competence of most contractors is satisfactory or very satisfactory. They feel much the same about inhouse researchers (75% satisfactory or very satisfactory) but a larger percentage considered their competence minimally acceptable (7%) or somewhat inadequate (7%). Comments on this item refer to contractors trying to cover too broad a spectrum with a limited staff. Senior researchers may be outstanding; however, what counts is the staff actually assigned to the contract. One problem that may arise with inhouse personnel is their reluctance to work outside their primary area of interest. Also administrative responsibilities interfere with research activities.

Managers are much less laudatory about the way in which research reports are written. Only 61% consider these satisfactory or very

satisfactory: 17% consider them only minimally acceptable and 16% consider their writing somewhat or completely inadequate. Writing is often poor, too long, too erudite, written for other professionals, does not communicate to the operational user.

The clearest trend one finds in manager responses is that pressure from higher management to justify research utility is increasing. 60% of managers felt that this pressure is very great, 40% say it is either moderate or great. What is not clear is how government defines research utility. The General Accounting Office defines it in terms of whether the ultimate user of the research report - as specified by the researcher himself - reports that he has done something or other with that report but this is a less than optimal criterion. Usefulness can also be defined (directly or indirectly) as related to implementation of the research product, but research implementation within government has been notoriously poor. In any event, the consequence of the push toward R&D utility may be to emphasize topics that have "grabber appeal", that are in other words immediately attractive even though of dubious value in the long range. Another consequence of the push to utility is the increasing length of time that managers must spend in writing justifications both before and after the research is initiated.

Pressure from executive and congressional levels for greater research utility suggests that they are somewhat skeptical about the usefulness of behavioral R&D. Nevertheless, 29% of laboratory managers felt that the usefulness of behavioral research in solving important methodological problems was great; 40% felt that it was moderate; 25%, slight and 3%, non-existent.

There is somewhat greater approbation of research utility for operational problems, 78% of managers reporting that they felt that utility was either moderate (58%), great (17%) or very great (3%). Unfortunately 20% still feel that this type of utility has been slight.

Government personnel have a somewhat less optimistic viewpoint than their contractor opposite numbers about the probability of applying standards of research rigor to the research situations in which they work. 42% of the contractors felt that it was usually possible to apply these standards; only 10% of the government managers felt this way. Whereas 30% of the contractors felt it was rarely possible to apply these standards to their research, 43% of the government managers felt this way. One

respondent pointed out that behavioral problems are generally multivariate and therefore only partially controllable. Another indicated that rigid application of rigorous research standards might well be self-defeating; there may be certain problems for which such standards are not necessary. Obviously, it is more possible to apply such standards to basic research conducted in a quasi-laboratory environment than to exploratory or developmental research.

If there are problems in utility it may be because HF does not really have the technology necessary to attack the problems given it by higher management. Only 50% of the government researchers felt that HF had the necessary technology; 24% felt that it did not, the remainder, 28% felt that it possessed that technology only for certain areas.

HF Practitioners

The HF practitioner is the behavioral specialist who, working in industry on some aspect of system development, must apply the research sponsored by government laboratories and performed by contract researchers. As one would expect, the percent of R&D performed in the companies in which practitioners work is heavily funded by government; the mean is 52%, but as before, there is high variability, some companies doing little or nothing, others doing almost all their work for the government. The HF group within the company spends 63% of its time supporting system development (whether or not funded by the government), 24% of its time on government-funded research contracts and 13% on a number of incidental activities.

Since the ultimate consumer of much behavioral research is the practitioner who attempts to apply it in development, one must ask whether this research as reported in journals like Human Factors, Ergonomics or the Journal of Applied Psychology is generally applicable to that system development. Over half (52%) of respondents moderately (44%) or strongly (8%) disagree that it is applicable. The remainder felt that the research does apply. Why we should get this bimodal distribution is not at all clear and one will have to perform further analyses to see if there is some factor that underlies the split. It is clear however that many practitioners do not believe they are getting the benefits of behavioral research. Sample comments are: narrow, theoretical, very little generalizability to system development. This is some sort of

judgment on the government which supports that research.

The relationship between the practitioner and the design engineer is important because their interaction is crucial to the application of HF data. 36% of practitioners moderately and 28% strongly disagree with the statement that design engineers on their own are capable of understanding HF inputs. However, some respondents pointed out that engineers usually understand these inputs but do not push for their incorporation in design. There are the usual large individual differences in designer practitioner relationships.

Nor do design engineers tend to solicit the assistance of practitioners. 76% of respondents agreed with this statement. Again, there are individual variations, special individuals and special circumstances but the armed neutrality between designer and practitioner seems the same as it was when it was described in 1967 (Meister and Farr, 1967). A key element in securing designer cooperation appears from respondents' comments to be a supportive management. A number of factors appear to explain the designer-practitioner relationship: the designer's wish to function with complete autonomy; his view of HF requirements as more constraints he must put up with; the HF group's reputation. It is helpful if the HF group has sign-off on man-machine interface drawings, but few groups have this sign-off.

The application of research data to system development may or may not be hindered by many practitioners' perception of HF as an art rather than a science. 60% either strongly (3%) or moderately (57%) agreed with that statement. As an art, HF application would seem to depend more on the special talents of the practitioner than on a formal body of principles and data. The attitude expressed seems characteristic of a primitive rather than a developed discipline. It is, of course, conceivable that the attitude (HF as art rather than science) results from the lack of behavioral principles and data, a lack which forces the practitioner to depend largely on his intuition.

By a very large majority (78%) practitioners feel that behavioral research data are generally inadequate to answer HF questions arising during system development. The studies available are apparently just enough different from the specific development situation to which they must be applied that their data cannot be accepted without reservation. In general,

the practitioner needs more specific detail than he can find in the literature. Some aspects such as controls or displays are well covered in the literature, whereas others, e.g. task design, are not.

Slightly more than half (57%) of practitioners feel that there is still considerable resistance on the part of designers to the inclusion of HF inputs in design. The positive side is that almost half (41%) do not agree with this notion. It may be that these responses suggest that things are improving somewhat, because in years past almost all practitioners would have given negative answers on this point. Some practitioners feel that if behavioral inputs are reasonable, engineers will accept them. Unfortunately some HF inputs are inadequate and this creates resistance to or rather avoidance of the inputs. Timing is all-important; inputs made after decisions have been reached by designers will be resisted.

This resistance may result in part from the fact that engineers may find HF inputs to design insufficiently precise and quantitative. 72% of the practitioners felt this to be the case. Some pointed out that HF data must be translated by practitioners into specific design terms or else the input is merely an additional burden to the engineer.

Not unexpectedly, almost all the practitioners (96%) feel that, left on their own, engineers would not incorporate HF considerations in design as effectively as would the practitioners. There is a considerable range of individual differences here, a few engineers being highly proficient in HF, others much less so. It is probable that if engineers handled behavioral problems on their own, obvious problems might be caught but not the more subtle ones. Design tends to be an adversary process between inputs from different disciplines and so motivation - or rather the lack of it - to incorporate behavioral inputs becomes critical. Moreover, the engineer tends to respond to immediate rather than to potential problems; behavioral inputs usually relate to the latter.

If HF is not more influential, is it because the necessary backup data are not available? 48% agree with this point; 51% do not. 44% of practitioners feel that human engineering handbooks and military standards provide sufficient information to handle the great majority of HF design situations encountered. Unfortunately, 54% disagree. 80% feel that the behavioral research performed under government sponsorship does not

sufficiently address system development problems. On the other hand, some practitioners point out that data are available but may not be used adequately; in other words, that the interpretation of research results is as important perhaps as the research itself.

Another factor that may underly the less than glowing impact of behavioral inputs on design is that the government does not effectively monitor development of the systems it funds and thus fails to ensure the company's application of human engineering standards to design. Behind this is the assumption that industry does not willingly make use of behavioral inputs. 80% of practitioners agree that the government does not monitor effectively and 88% feel that the HF effort in system development is not adequately funded.

It is interesting that there is a general feeling on the part of practitioners that strong government monitoring of HF efforts is an ally in their struggle against recalcitrant company management. The assumption is made that since industry takes the position that the customer is always right, if government insists on a strong HF effort, industry will give it to them.

64% of practitioners are gloomy about the influence of the HF discipline on overall system development; they feel it is minimal. The same majority (69%) feel other priorities such as cost or reliability seriously diminish the influence of HF inputs on design. These priorities may be more management than engineering-directed. On the other hand, certain priorities like performance requirements should indeed supersede everything else.

With all this there is reason for hope. 76% of practitioners feel that over the years designers have shown increasing appreciation of the importance of HF in design. Again, they point out that there is considerable variation among engineers, the older ones tending to be more conservative.

Conclusions

The results described previously in most cases speak for themselves. However, the following points should be made:

1. The influence of government in behavioral R&D is overwhelming - it is perhaps the one most significant element in the picture.
2. There is a discontinuity between those who request research and those who carry it out. Selection of topics (at a rather gross level only, of course) is performed by bureaucrats who are not specialists; more specific plans are made by laboratory management which almost never does any of the research itself; the work is performed either inhouse or by contract; and if the latter, the contractor has little or no influence on what is done, although he may have some on how it is done. Thus, there is authority without responsibility and responsibility without authority.
3. Dept. of Defense agencies break out into: a. those largely concerned with basic research like ONR or ARPA; b. those concerned mostly with applied research like AFHRL; c. those that do most of their work inhouse; d. those that contract out most of their work or that have a roughly 50-50 mix. The bi-modal distributions found in the questionnaire responses are largely due to mission differences which, in turn, influence the manner in which behavioral R&D is performed and the results utilized. Among practitioners we find those whose work is exclusively or largely on government supported development, those who work on support of inhouse company projects, etc.
4. Despite these differences, we find a substantial dissatisfaction in all three categories - contractors, laboratory personnel, practitioners - with the way in which behavioral R&D is performed and the results that it provides. There is often a lack of clarity about what the government wants when it contracts out and the time frame imposed on contractors is too short in many cases to do an effective job. Results are only rarely implemented.
5. It is clear that there is continuing and increasing pressure to justify the utility of behavioral R&D. While this may not be unfortunate in and of itself, it does lead to a number of unfortunate results: faddism; impatience with studies whose effects are slow to emerge; unwillingness to invest research resources where results are risky.
6. If, as many believe, a major goal of behavioral R&D is to provide the data that will help optimize new system development, that goal, in the opinion of practitioners, is apparently not being wholly

fulfilled. This situation is exacerbated by the continuing (although dissipating) resistance of design engineers to the utilization of behavioral inputs. The lack of adequate data and governmental inadequacy in monitoring the company's HF effort contribute to this resistance.

7. It is, of course, only an assumption that the respondents' perceptions are valid. However, if one accepts this assumption, then clearly the situation needs improvement. What is needed initially and as a first priority is to ensure that knowledgeable and disinterested behavioral specialists make some sort of input to the governmental managerial decision process which leads to the selection of R&D topics. If the point is advanced that executive agencies do have behavioral specialists to advise them, then one can say only that either these specialists are inadequate (which is unlikely) or they do not have sufficient influence with their managers to make meaningful inputs to the R&D decision process.
8. Some way needs to be developed to permit contractors to make a contribution to the selection of research development topics. 5-year plans (which they might critique) are sometimes developed by governmental agencies, but these tend to be vague. Nevertheless, a mechanism should be developed to permit some contractor representatives to express their attitudes toward projected research. Similarly, practitioner representatives should also be permitted an input to these plans from their special standpoint.
9. These recommendations sound rather vague and undoubtedly they are - a problem involving three segments of the R&D community cannot be solved with a wave of the pen. It would seem worthwhile however to call representatives of the behavioral R&D community together periodically for a sort of congress to examine projected R&D plans, to critique them and to offer suggestions.

References

Meister, D. and Farr, D. The Utilization of Human Factors Information by Designers. Human Factors, 1967.

Note: The opinions expressed in this paper are those of the author alone and not of the U.S. Navy.